

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Patent Application of:

Masao OOTA

Application No.: (Unassigned)

Group Art Unit:

Filed: (Concurrently)

Examiner:

For: METHOD FOR STORING DATA USING GLOBALLY DISTRIBUTED STORAGE SYSTEM, AND PROGRAM AND STORAGE MEDIUM FOR ALLOWING COMPUTER TO REALIZE THE METHOD, AND CONTROL APPARATUS IN GLOBALLY DISTRIBUTED STORAGE SYSTEM

**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN  
APPLICATION IN ACCORDANCE  
WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55**

Commissioner for Patents  
PO Box 1450  
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No(s). 2002-286528

Filed: September 30, 2002

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: 9/22/03

By: Richard A. Gollhofer  
Richard A. Gollhofer  
Registration No. 31,106

1201 New York Ave, N.W., Suite 700  
Washington, D.C. 20005  
Telephone: (202) 434-1500  
Facsimile: (202) 434-1501

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年 9月30日

出 願 番 号

Application Number:

特願2002-286528

[ST.10/C]:

[JP2002-286528]

出 願 人

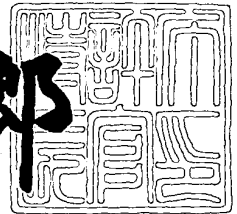
Applicant(s):

富士通株式会社

2003年 1月24日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

太田信一郎



出証番号 出証特2003-3001442

JAPAN PATENT OFFICE

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: September 30, 2002

Application Number: Patent Application No. 2002-286528  
[ST.10/C]: [JP2002-286528]

Applicant(s): FUJITSU LIMITED

January 24, 2003

Commissioner,

Japan Patent Office Shinichiro OTA

Certificate No. P2003-3001442

【書類名】 特許願

【整理番号】 0252064

【提出日】 平成14年 9月30日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 3/06

【発明の名称】 広域分散ストレージシステムを利用したデータ格納方法、その方法をコンピュータに実現させるプログラム、記録媒体、及び広域分散ストレージシステムにおける制御装置

【請求項の数】 5

【発明者】

【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

【氏名】 太田 昌男

【特許出願人】

【識別番号】 000005223

【氏名又は名称】 富士通株式会社

【代理人】

【識別番号】 100074099

【住所又は居所】 東京都千代田区二番町8番地20 二番町ビル3F

【弁理士】

【氏名又は名称】 大菅 義之

【電話番号】 03-3238-0031

【選任した代理人】

【識別番号】 100067987

【住所又は居所】 神奈川県横浜市鶴見区北寺尾7-25-28-503

【弁理士】

【氏名又は名称】 久木元 彰

【電話番号】 045-573-3683

【手数料の表示】

【予納台帳番号】 012542

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9705047

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 広域分散ストレージシステムを利用したデータ格納方法、その方法をコンピュータに実現させるプログラム、記録媒体、及び広域分散ストレージシステムにおける制御装置

【特許請求の範囲】

【請求項 1】 コンピュータが、データを冗長化して複数のボリウムに分割し、各ボリウムを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法であって、

帯域幅、通信コスト及び書き込みを依頼するノードからストレージまでの経路の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを特徴とするデータ格納方法。

【請求項 2】 前記データを読み込む際には、各ストレージについて、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、

前記利用優先度に基づいて冗長化部分を含まないボリウムとして、前記複数のボリウムのうちいずれのボリウムを各ストレージから読み出すか決定する、

ことを更に含むことを特徴とする請求項 1 に記載のデータ格納方法。

【請求項 3】 前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第 1 のボリウムの複製を格納する、

ことを更に含むことを特徴とする請求項 1 に記載のデータ格納方法。

【請求項 4】 ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御をコンピュータに行われるコンピュータ・プログラムであって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせることを特徴とするコンピュータ・プログラム。

【請求項 5】 ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリュームに分割し、各ボリュームを複数のストレージに分散して格納する制御を行う制御装置であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する経路管理手段と、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択するストレージセット管理手段と、

を備えることを特徴とする制御装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、システム内の格納装置を多重化することにより、データの冗長性や格納装置の性能の改善を図る技術、つまり R A I D (Redundant Array of Inexpensive Disks) に係わる技術に関する。

【0002】

【従来の技術】

従来、R A I Dを用いて、1つのデータを複数に分割し、分割されたデータを複数のストレージへ分散して格納することにより、システムの耐障害性（フォールトトレランス性）を向上させることが行われている。R A I Dには、レベル0からレベル6までの7レベルがあり、さらに、複数のR A I Dのレベルを組み合わせたレベルや共通化されていない独自レベルもある。このうち、レベル5は、データを複数に分割し、それらの分割されたデータの各々にパリティデータを付加し、それぞれの分割されたデータを複数のストレージに分散させて格納する。レベル5は、1台のネットワーク端末もしくは処理サーバと、ファイルサーバの

ように、比較的密な関係を持った装置間で採用されることが好ましい。

【0003】

図33に、RAIDを採用したシステムの構成図を示す。まず、図33において、ルータR（中継器）を用いて、ストレージサービスセンタSC、バックアップセンタBC、ミラーセンタMC及び利用者（ネットワーク端末あるいは処理サーバ）が接続され、これによって広域分散ストレージシステムが形成されている。

【0004】

以下、利用者が本社と支社とに分かれており、本社の利用者UHがデータを広域分散ストレージシステムにデータを格納し、そのデータを支社の利用者UBが利用すると仮定して、この広域分散ストレージシステムにおいて行われる処理について説明する。

【0005】

まず、本社の利用者UHは、格納したいデータをストレージサービスセンタSC内のストレージに格納させる。ストレージサービスセンタSCは、データを複製し、複製されたデータをバックアップセンタBC内のストレージに格納させる。なお、災害等によってストレージサービスセンタSCとバックアップセンタBCの双方が損害を受ける可能性を低くするために、ストレージサービスセンタSCとバックアップセンタBCとの物理的な距離は遠隔であることが望ましい。

【0006】

さらに、支社の利用者UBがデータをストレージから読み出す際のレスポンスを改善するために、ストレージサービスセンタSCはデータを複製し、複製されたデータを、支社に最寄りの接続地点となっているミラーセンタMC内のストレージに格納させる、又は、支社の利用者UBからストレージサービスセンタSCまでの回線に割り当てる帯域を広くする。なお、バックアップセンタBCがミラーセンタMCを兼ねることとしてもよい。

【0007】

また、RAIDに関する技術として、データをセグメントに分割し、そのセグメントごとに複数のストレージにランダムに分散格納する、つまり、ストライピ



ング先となるストレージをランダムに決定する第1の発明がある。第1の発明により、一次ストレージが故障した場合その負荷全体が二次バックアップストレージにかかるという問題、及びコンボイ効果の確度が高くなるという問題を解決することが可能となる（例えば、特許文献1参照）。

【0008】

また、さらなるRAIDに関する技術として、ストレージに格納されたデータをミラーリングする際に、そのデータを複数に分割し、分割されたデータを複数のストレージに分散して格納する第2の発明がある。第2の発明により、元となったストレージに障害が生じた場合でも、複数のストレージに分散格納されていたデータを読み出して、これらのデータを用いて、元となったストレージに格納されていたデータを復元することが可能となる（例えば、特許文献2参照）。

【0009】

また、さらなるRAIDに関する技術として、バッファに格納されたデータをストレージに書き出す際に、送出先となるストレージが同一であるデータが複数存在する場合、それらのデータを1つの複合パケットにまとめて送出する第3の発明がある。第3の発明により、RAIDのI/Oスループット性能を向上させることが可能となる（例えば、特許文献3参照）。

【0010】

【特許文献1】

特表2002-500393号公報（段落0005から段落0007、  
図1）

【0011】

【特許文献2】

特開平9-171479号公報（段落0018から段落0020、段落  
031から段落0034、図1）

【0012】

【特許文献3】

特開平10-333836号公報（段落0024から段落0027、図  
3）

【 0 0 1 3 】

【発明が解決しようとする課題】

しかし、上記の図 3 3 に示した従来に係わる広域ネットワークシステムには、以下に挙げる問題があった。

【 0 0 1 4 】

1) バックアップセンタ及び・又はミラーセンタに、ストレージサービスセンタ SC と同容量のストレージを備えることが必要であるため、システムが高コストとなる。

【 0 0 1 5 】

2) バックアップ又はミラーリングを行う場合、そのために回線を利用することとなり効率的ではない。

3) 利用者が利用するネットワーク端末或いは処理サーバが、マルチホーミングされている場合であっても、それによって利用可能となっている複数の回線を効率的に使用できない。

【 0 0 1 6 】

4) 各センタ SC、BC 又は MC 内のストレージ等が盗難された場合、ストレージに格納されていたデータが損害を受けるため、セキュリティがよくない。

また、上記に記載の第 1 の発明から第 3 の発明においても、上述の 1) から 4) の問題が解決されていない。

【 0 0 1 7 】

以上の問題に鑑み、データの冗長化に要するストレージ容量を低減しつつも、データのセキュリティを向上させ、且つ、回線を効率的に利用することが可能な RAID を提供することが、本発明が解決しようとする課題である。

【 0 0 1 8 】

【課題を解決するための手段】

上記問題を解決するために、本発明の 1 態様によれば、コンピュータが、データを冗長化して複数のボリュームに分割し、各ボリュームを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法において、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距

離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択することを含む。

#### 【0019】

データを複数のボリュームに分割して複数のストレージに分散して格納することによりデータのセキュリティを向上させつつ、さらに、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離を考慮してそのノードから見て最適なストレージを選択することにより、回線効率と災害時のデータの安全性も向上させることが可能となる。

#### 【0020】

上記方法における、前記評価値の算出の際に、更に、前記書き込みを依頼するノードから各ストレージまでのホップ数を考慮する事としても良い。ホップ数が高い場合は、回線効率が低下するからである。

#### 【0021】

また、上記方法において、前記システムの利用者に対して、前記ストレージセットを仮想的な1つのストレージとして提供することを更に含むこととしてもよい。これにより、データの分散格納によって、利用者にとって操作が複雑化することを避けることが可能となる。

#### 【0022】

また、上記方法において、前記データを前記ストレージセットから読み込む際には、前記ストレージセットに書きこまれた前記複数のボリュームのうち冗長化部分を含まないボリュームを各ストレージから読み出し、前記読み出されたボリュームを用いて前記データを復元する、ことを更に含むこととしてもよい。これにより、使用する回線帯域を抑制することが可能となる。

#### 【0023】

また、上記方法において、前記データを読み込む際には、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、前記利用優先度に基づいて、冗長化部分を含まないボリュームとして、前記複数のボリュームのうち

いずれのボリウムを各ストレージから読み出すか決定することを更に含むこととしてもよい。例えば、データを3データ+1パリティで冗長化して4つのボリウムに分割している場合、冗長化部分を含まないボリウムとして、3つのボリウムを任意に選択することができる。この選択の際に、帯域幅及びコストを考慮する事により、回線の利用効率を向上させることが可能となる。

## 【 0 0 2 4 】

また、上記方法において、前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第1のボリウムの複製を格納することを更に含むこととしてもよい。この複製はバックアップとして用いることができる。従来、バックアップとしてデータの複製を備えていたため、バックアップを備えるためには元データの2倍の容量が必要であった。しかし、この態様によれば、バックアップのために必要なストレージ容量は、ただか1ボリウム分の容量となる。これにより、ストレージの使用効率を向上させることが可能となる。

## 【 0 0 2 5 】

また、上記方法において、前記第1のボリウムの複製を作成する際に、前記評価値に基づいて、前記第1のボリウムを格納するストレージから前記第1のボリウムを複製するのか、前記複数のボリウムのうちの前記第1のボリウム以外のボリウムから冗長を利用して前記第1のボリウムを再現するのか、2つの作成方法のうちのいずれかを選択することを更に含むこととしてもよい。ここで、この選択において前記評価値を考慮することとしてもよい。

## 【 0 0 2 6 】

また、上記方法において、同一のボリウムを格納すべき複数のストレージに対して、マルチキャストでボリウムを書き込むこととしてもよい。これにより、同じ内容をもつパケットを複数回にわたり送信することを回避する。

## 【 0 0 2 7 】

また、上記方法において、前記第1のボリウムの複製をストレージに書き込む際に、多数回に分けて書き込み処理を行うこととしてもよい。多数回に分けることにより、一度につき回線にかかる負荷を軽減する事が可能となる。

## 【 0 0 2 8 】

また、上記方法において、前記ストレージセットのうちの第1のストレージに障害が発生した場合、前記ストレージセットのうちの他のストレージへの書き込みを制限することを更に含むこととしてもよい。例えば、第1のストレージの復旧前に他のストレージ内のボリュームが更新されてしまう事がありうるが、これにより、障害が発生したストレージの復旧後、システム内に異なるバージョンのボリュームが並存する事を防止する。

## 【0029】

また、上記方法において、前記ストレージセットのうち第3のストレージに障害が発生した場合、前記評価値に基づいて、前記ストレージセットとして選択されているストレージ以外の第4のストレージを、前記第3のストレージの代わりを選択することとしてもよい。これにより、障害が発生したストレージの代わりとして最適なストレージを選択することが可能となる。

## 【0030】

また、上記方法において、前記ストレージセットの選択後、一定のタイミングで、各ノードにおけるストレージセットを再選択し、再選択の結果、どのノードからも利用されていないボリュームがあった場合、該ボリュームをストレージから削除することを更に含むこととしてもよい。ここで、一定のタイミングとは、前回の選択から一定期間後、又はボリュームの状態が変更される毎である。システムの使用状況が変化するタイミングにおいて、ボリュームの使用状況に基づいて不要なボリュームを削除する事により、ストレージの利用効率を向上させることが可能となる。

## 【0031】

また、上記方法において、前記データを読み込んだ後に、前記データを一定期間、任意の1つのストレージ内に一時格納し、前記一定期間内にデータの読み出しを行う際には、一時格納されたデータを前記1つのストレージから読み出すことを更に含むこととしてもよい。このようなキャッシュ機能を備える事により、データの読み出しレスポンスを向上させることが可能となる。

## 【0032】

また、上記方法において、一定期間内に書き込み要求されたデータを一時記憶

領域に保持し、前記一定期間経過後に一時格納領域からデータを取り出し、データを複数のボリュームに分割し、該複数のボリュームを前記ストレージセットに書き込むことを更に含むこととしてもよい。これにより、書き込み要求を出すノードから他のストレージへボリュームを転送する回数を低減する事が可能となるため、トラフィックの効率を上げることが可能となる。

#### 【0033】

また、上記方法において、前記複数のストレージに前記複数のボリュームを書き込む際に、前記書き込みを依頼するノードは、書き込みが終了するまで前記複数のストレージへの書き込み処理を禁止することを更に含むこととしてもよい。ここで、同一のボリュームを格納すべき複数のストレージがある場合、それらのストレージの中から、1つのストレージを代表ストレージとして決定し、前記複数のストレージへの書き込み処理の禁止において、前記代表ストレージへの書き込み処理の禁止は、前記書き込みを依頼するノードによって行われ、前記代表ストレージ以外のストレージへの書き込み処理の禁止は、前記代表ストレージによって行われることとしてもよい。また、前記代表ストレージは、原本となるボリュームを格納すべきストレージであることとしてもよい。

#### 【0034】

また、上記方法に含まれる手順を含む制御をコンピュータに行わせるコンピュータ・プログラムも、コンピュータによって上記コンピュータ・プログラムを実行させる事によって、上記方法と同様の作用・効果が得られるため、上記課題を解決することが可能である。

#### 【0035】

また、上記コンピュータ・プログラムを記録したコンピュータ読み取り可能な記録媒体から、そのプログラムをコンピュータに読み出させて実行させることによっても、上記課題を解決することができる。

#### 【0036】

また、上記データ格納方法において行われる手順と同様の処理を行う、ネットワークを介して分散配置されたストレージを備えるシステムにデータを分散格納させる制御を行う制御装置によっても、上記データ格納方法と同様の作用・効果

が得られるため、上記課題を解決することが可能である。

【 0 0 3 7 】

【発明の実施の形態】

以下、本発明の実施の形態について図面を用いて説明する。なお、同じ装置等には同じ参照番号をつけ、説明を省略する。なお、以下の説明において、「ノードに備えられるストレージ」をいう際に「ノード」という場合がある。これは、文が長くなるために意味が分かりにくくなる事を防ぐためである。例えば、「ノードにボリウムを格納する」という表現は、「ノードに備えられるストレージにボリウムを格納する」ということを意味する。

【 0 0 3 8 】

本発明は、データにパリティを付加し、これらを複数のストレージに分散格納する技術、例えば R A I D 5 等の技術を前提とする。図 1 に、本発明の各実施形態に係わる広域分散ストレージシステムの構成を示す。図 1 に示すように、広域分散ストレージシステムにおいて、ネットワークを介して複数のノードが接続されている。ノード間で通信されるデータは、ルータ R で中継される。各ノードは、ストレージ S 及び制御装置 C を備える。

【 0 0 3 9 】

利用者本社や利用者支店等に備えられた端末の利用者は、広域分散ストレージシステムにアクセスし、ストレージ S にデータを格納させたり、ストレージ S からデータを読み出したり等を行う。

【 0 0 4 0 】

制御装置 C は、端末からデータをストレージに格納するよう指示された場合、格納すべきデータのデータブロック単位（読み出し／書き込みの単位）に E C C （Error Check and Correct）／パリティを付し、複数のストレージ S にデータを分散させて格納させる。以下、分割されてパリティを付加されたデータをボリウムという。

【 0 0 4 1 】

端末からストレージ内に格納されたデータを読み出すよう指示された際には、制御装置 C は、複数のストレージ S から分散されて格納されているデータ、つま

りボリウムを読み出して、データを復元して端末に送信する。

【 0 0 4 2 】

データの格納及び読み出しの際、データの分散格納及び復元を制御装置 C が行うため、端末の利用者は、データが分散されていることを意識することなく、1 つの仮想ディスクにデータを格納し、その仮想ディスクからデータを読み出す際と同様に、データを分散格納したり復元したりすることができる。

【 0 0 4 3 】

また、ボリウムを読み出してデータを復元する際に、制御装置 C は、そのデータを構成する全てのボリウムを複数のストレージから読み出してそのデータを復元する。或いは、制御装置 C は、冗長化された分のボリウムを除いたボリウムを複数のストレージ S から読み出してそのデータを復元する事としても良い。この場合、ネットワークにかかる負荷を低減する事ができる。より具体的には、2 データ + 1 パリティで冗長化されて分割されたデータを復元する場合、制御装置 C は、3 つのボリウムのうち 2 つのボリウムを読み出してデータを復元する。

【 0 0 4 4 】

図 2 に、制御装置 C の構成を示す。図 2 に示すように、制御装置 C は、ユーザインタフェース（以下、ユーザ I F）（受信側）1、ユーザ I F（送信側）2、データ変換部 3、パケット生成部 4、制御部 5、データ組立部 6、パケット解析部 7、ストレージインタフェース（以下、ストレージ I F）8、ネットワークインタフェース（以下、ネットワーク I F）（送信側）9 及びネットワーク I F（受信側）10 を備える。

【 0 0 4 5 】

ユーザ I F（受信側）1 は、利用者からストレージ S へアクセスするパケットを受信し、制御情報を制御部 5 に、データをデータ変換部 3 へ振り分ける。

データ変換部 3 は、データをデータブロックに分割し、各ブロックにパリティを付加する。

【 0 0 4 6 】

パケット生成部 4 は、広域ネットワークに送信するために、ブロック単位に分割されたデータ又は制御情報をパケット化する。



ネットワーク I F（送信側）9 は、パケット生成部 4 によって生成されたパケットをネットワークに送信する。

【 0 0 4 7 】

ネットワーク I F（受信側）1 0 は、広域ネットワークからデータ又は制御情報を受信する。

パケット解析部 7 は、ネットワーク I F（受信側）1 0 から出力されたパケットを解析し、ストレージ S からのデータ読み出し又はストレージ S へのデータ書き込み処理を行う。

【 0 0 4 8 】

データ組立部 6 は、ストレージ S から読み出された信号を組み立てて、利用者からのデータアクセス指示に対して、制御情報を含む適正なパケットを生成する。

【 0 0 4 9 】

制御部 5 は、利用者からのアクセスに基づいて、ストレージ S 及びデータの管理及び送受信パケットを処理する。

ユーザ I F（送信側）2 は、利用者に、データ組立部 6 によって組み立てられたパケットを送信する。

【 0 0 5 0 】

次に、図 3 に、制御装置 C の詳細構成図を示す。以下、図 3 に示す詳細構成図にそって、データ変換部 3、パケット生成部 4、制御部 5 及びデータ組立部 6 の動作について詳しく説明する。

【 0 0 5 1 】

上記データ変換部 3 は、パケット解析部 3 0 1、データ分割部 3 0 2 及びパリティ計算部 3 0 3 を備える。パケット解析部 3 0 1 は、受信されたパケットを解析し、そのパケットからデータを取得する。データ分割部 3 0 2 は、データをデータブロック単位に分割する、パリティ計算部 3 0 3 は、パリティを計算し、データブロックに付加する。

【 0 0 5 2 】

上記パケット生成部 4 は、データ管理情報付加部 4 0 1、制御／経路情報負荷

部 4 0 2、データ転送部 4 0 3 及び転送パケット構築部 4 0 4 を備える。データ管理情報付加部 4 0 1 は、データブロックに制御部 5 からの出力されたデータ管理情報を付加する。なお、データ管理情報は、ストレージセット構成情報（後述）等であり、そのパケットに基づいて行われる処理の内容によって異なる。各処理において送信されるデータ管理情報については、後述する。

#### 【 0 0 5 3 】

制御／経路情報付加部 4 0 2 は、データブロックに、制御情報や経路情報を付加する。なお、経路情報は、そのデータブロックの宛先となるノードまでの経路及びその経路の評価値を示す情報であり、制御部 5 によって生成される。制御情報は制御の内容、例えばデータの書き込みであるのか、読み込みであるのか、ストレージへの書き込みを制御するのか等を示す情報であり、制御部 5 によって生成される。

#### 【 0 0 5 4 】

データ転送部 4 0 3 は、データ管理情報及び制御／経路情報が付加されたデータパケット又は、制御部 5 から出力された制御パケットを転送する。転送する際に、データ転送部 4 0 3 は、パケットの宛先となるノードのアドレス、例えば I P (Internet Protocol) アドレス等をパケットに付加する。このアドレスは、制御部 5 から出力される。なお、制御／経路情報に基づいて、データが、そのノード内のストレージに書き込まれるべきデータ（ローカルデータ）であると判定された場合、データ転送部 4 0 3 は、そのデータをパケット解析部 7 に出力する。

#### 【 0 0 5 5 】

転送パケット構築部 4 0 4 は、ストレージ I F 8 を介してストレージ S から読み出されたデータを他のノードの制御装置 C に転送するための転送パケットを構築し、データ転送部 4 0 3 に出力する。パケットを構築する際に、転送パケット構築部 4 0 4 は、上記のデータ管理情報付加部 4 0 1 及び制御経路情報付加部 4 0 2 と同様の処理を行う。

#### 【 0 0 5 6 】

上記制御部 5 は、ストレージ制御部 5 0 1、制御パケット生成部 5 0 2、ネッ

トワーク制御部 5 0 3、経路管理部 5 0 4、ストレージセット管理部 5 0 5、ローカルボリューム管理部 5 0 6、経路評価テーブル 5 0 7、ストレージ評価テーブル 5 0 8、ストレージセット管理テーブル 5 0 9、アクセス管理テーブル 5 1 0 及びローカルボリューム管理テーブル 5 1 1 を備える。ストレージ制御部 5 0 1 は、パケット解析部 3 0 1 から出力された制御情報に基づいて、ストレージ S へのデータ書き込み、読み出し及びロック等を制御する。また、ストレージ制御部 5 0 1 は、制御パケット生成部 5 0 2、ネットワーク制御部 5 0 3、経路管理部 5 0 4、ストレージセット管理部 5 0 5 及びローカルボリューム管理部 5 0 6 の動作の連携も制御する。

#### 【 0 0 5 7 】

制御パケット生成部 5 0 2 は、データ書き込み、データ読み込み、ストレージ S のロック等の制御の内容を示す制御パケットを生成する。この制御パケットは他のストレージへ送信される。ネットワーク制御部 5 0 3 は、経路管理部 5 0 4 からの出力に基づいて、パケットの宛先となるノードを示す経路情報及びそのノードのアドレス等を生成する。なお、ノードのアドレスは、不図示のアドレステーブルに登録されているとする。アドレステーブルについては自明であるため、ここでは説明しない。

#### 【 0 0 5 8 】

経路管理部 5 0 4 は、経路評価テーブル 5 0 7 及びストレージ評価テーブル 5 0 8 に格納された情報に基づいて、ブロック単位に分割されたデータの宛先つまり、データの格納先又は転送先等を決定する。ストレージセット管理部 5 0 5 は、ストレージセット管理テーブル 5 0 9 を用いて、データを構成する複数のボリュームを管理する。また、ストレージセット管理部 5 0 5 は、各ノードのストレージ S に格納されたボリュームが更新される際に、アクセス管理テーブル 5 1 0 を用いて各ストレージへのアクセスを管理する。ローカルボリューム管理部 5 0 6 は、ローカルボリューム管理テーブル 5 1 1 を用いて、ローカルストレージの利用状況を管理する。各テーブルの構造について詳しくは後述する。

#### 【 0 0 5 9 】

上記データ組立部 6 は、パケット構築部 6 0 1、データ組立部 6 0 2 及びパリ

ティ計算部 6 0 3 を備える。パリティ計算部 6 0 3 は、パリティを計算する。データ組立部 6 は、パリティ及びボリウムがどのボリウムであるのかを示すボリウム番号に基づいて、ローカルストレージ S から読み出されたボリウムのデータ（ボリウムデータ）やその他のノードから受信したパケット内のボリウムデータを用いて、分割される前のデータを復元する。なお、ボリウム番号は、ボリウムデータに付されている。パケット構築部 6 0 1 は、データアクセス指示を出した利用者に、復元されたデータを送信するためにパケットを生成する。

#### 【 0 0 6 0 】

次に、図 4 に、広域分散ストレージシステムの具体的な構成例を示す。以下、図 4 に示す具体的な構成を用いて各テーブルの構成や制御装置の動作等を説明するが、これは具体的に説明するために図 4 に示すような構成を仮定しているのであり、ストレージシステムの構成を限定する趣旨ではない。

#### 【 0 0 6 1 】

図 4 に示すように、広域ネットワークを介して、ノード A からノード G から接続されている。各ノードには制御装置 C 及びストレージ S が備えられている。ノード A とノード B の間（以下、区間 A - B）、ノード E とノード F の間（以下、区間 E - F）及び「ノード F とノード G の間（以下、区間 F - G）の帯域幅は、1 5 0 M b p s である。ノード B とノード C の間（以下、区間 B - C）、ノード C とノード D の間（以下、区間 C - D）及びノード D とノード E の間（以下、区間 D - E）の帯域幅は、5 0 M p b s である。ノード G とノード A の間（以下、区間 G - A）の帯域幅は 1 G b p s である。ノード B とノード E の間（以下、区間 B - E）の帯域幅は、6 0 0 M b p s である。

#### 【 0 0 6 2 】

以下、図 5 から 9 を用いて、制御装置 C に備えられるテーブルの構造について説明する。まず、図 5 を用いて、経路評価テーブル 5 0 7 の構造について説明する。経路評価テーブル 5 0 7 は、広域分散ストレージシステムを構成する各ノードについての経路評価情報を格納する。経路評価テーブル 5 0 7 は、各ノード間を接続する経路の優位性を評価する際に参照される。図 5 に示すように、経路評価情報は、区間を識別する記号、その区間での帯域幅、その区間での通信コスト

、その区間の物理的距離（ディスタンス）、及びストレージの利用優先度等を項目として含む。また、経路評価情報は、さらに区間の利用優先度を更に含むこととしても良い。なお、ローカルノード、つまり、その経路評価テーブル507を備える制御装置Cが属するノードについては、ネットワークを介して通信する必要がないため、帯域幅、コスト及びディスタンスは空である。

## 【0063】

帯域幅、通信コスト及び物理的距離は、広域分散ストレージシステムの構成に基づいて決定され、どのノードに備えられた制御装置であっても、基本的に同じ値となる。ストレージの利用優先度及び区間の利用優先度は、制御装置Cによって計算される値である。利用優先度は、帯域幅及びコストに加えて、災害等が発生した際のデータバックアップの安全性を評価するためにディスタンスに基づいて決定される。

## 【0064】

ストレージ利用優先度の計算式は以下のとおりである。

ストレージ利用優先度

$$= (\text{帯域幅} \times A) \div (\text{コスト} \times B) + (\text{ディスタンス} \times C)$$

なお、A、B及びCは、重み付け定数である。以下の説明では、例として、A、B及びCをそれぞれ、2、1及び0.1と仮定する。各重み付け定数は、システムにおいて優先すべき性能、例えば通信速度を優先するのか、又はコストを優先するのか等を考慮して、変更することとしても良い。

## 【0065】

図5は、図4にシステムのノードAについての経路評価テーブル507を示している。以下、具体的に、区間A-Bについてストレージの利用優先度を算出する。

## 【0066】

区間A-Bについてのストレージの利用優先度

$$= (150 \times 2) \div (100 \times 1) + (80 \times 0.1) = 11$$

従って、図5に示す経路評価テーブル507において、区間A-Bについてのストレージの利用優先度として、「11」が格納されている。

## 【 0 0 6 7 】

なお、区間の利用優先度は、帯域幅÷コストを正規化した値である。

次に、図 6 を用いて、ストレージ評価テーブル 5 0 8 の構造について説明する。ストレージ評価テーブル 5 0 8 は、広域分散ストレージシステムを構成する各ノードに備えられたストレージについてのストレージ評価情報を格納する。ストレージ評価テーブル 5 0 8 は、ストレージセットの作成及び追加等を行う際に、どのノードのストレージを使用すべきか決定する際に参照される。図 6 に示すように、ストレージ評価情報は、ノードを識別する記号、ローカルノードからそのノードへの経路、ストレージ評価値及びホップ数等を項目として含む。ここで、ローカルノードとは、そのストレージ評価テーブル 5 0 8 を備える制御装置 C のノードをいう。ホップ数とは、あるノードに到達するまでの経路に介在するノード数をいう。ストレージ評価値は、制御装置 C によって計算され、その計算式は以下の通りである。

## 【 0 0 6 8 】

ストレージ評価値＝

$$\Sigma \{ (\text{経路上のノードのストレージの利用優先度}) \times (\text{重み付け定数}) \} \\ \div (\text{その経路の最終ノードまでのホップ数})$$

ここで、重み付け定数をホップ数の逆数とすると、ストレージ評価値の計算式は以下のようなになる。

## 【 0 0 6 9 】

ストレージ評価値

$$= \Sigma \{ (\text{経路上のノードのストレージの利用優先度}) \div (\text{そのノードまでのホップ数}) \} \div (\text{その経路の最終ノードまでのホップ数})$$

図 6 は、図 4 に示すシステム内のノード A に備えられるストレージ評価テーブル 5 0 8 を例示している。以下、例として、具体的に、ノード A から見たノード B のストレージの評価値及びノード C のストレージの評価値を算出する。

## 【 0 0 7 0 】

ノード B のストレージ評価値

$$= (\text{区間 A - B のストレージの利用優先度} \div 1) \div 1$$

$$= 11$$

ノードCのストレージ評価値

$$= \{ ( \text{区間A-Bのストレージの利用優先度} \div 1 ) + ( \text{区間B-Cのストレージの利用優先度} \div 2 ) \} \div 2$$

$$= \{ 11 \div 1 + 17 \div 2 \} \div 2$$

$$= 9.75$$

また、ストレージ評価情報は、さらに、経路評価値を項目として含むこととしても良い。経路評価値は、最終ノードに到達するまでに経由する区間の利用優先度の和をホップ数で割算する事により計算される。

#### 【0071】

次に、図7を用いて、ストレージセット管理テーブル509の構造について説明する。ストレージセット管理テーブル509は、それぞれのストレージセットに関する情報を管理するためのテーブルである。ストレージセット管理テーブル509は、ストレージセット構成情報を格納する。ストレージセット構成情報は、ストレージセットを識別するストレージセット番号と、その番号に対応するストレージセットに関する情報、つまりプロパティを含む。

#### 【0072】

ここで、ストレージセットとは、システム全体から見ると、データを分割されたことにより得られた複数のボリュームを分散格納するストレージをいう。しかし、後述のように、通常、各ノードにおいて、ボリュームを分散格納するストレージの全てを使用（書き込み、読み出し等行う）するのではなく、使用するストレージはこれらのうちの少なくとも一部である。各ノードで使用するストレージは、後述のプロパティに含まれる使用状況情報によって管理される。使用するストレージ以外は、バックアップ等の機能を果たす。従って、各ノードから見ると、ストレージセットとは、使用状況情報によって使用が許可されているストレージである。

#### 【0073】

ストレージセット番号は、ストレージセットを識別するために用いられるが、データを識別する情報としても利用される。例えば、システムの利用者は、読み

出したいデータを特定するためにストレージセット番号を用いる。これは、ストレージセットは、利用者に対して1つの仮想的なストレージとして提供されるからである。

#### 【0074】

プロパティには、広域分散ストレージシステム全体のプロパティと、各ノードのプロパティとがある。システム全体のプロパティは、データが分割されたノード数及びストレージの状態（良好であるか、異常があるのか等）を示す情報を含む。なお、データが分割されたノード数は、読み出しの場合と書き込みの場合のそれぞれについて格納される。なお、図7において、ストレージの状態として「G」が格納されている場合、状態が「良好」であり、「R」が格納されている場合、状態が「異常」である。

#### 【0075】

ノードのプロパティは、そのノードのストレージがストレージセット中のどのボリュームを格納するのかを示すボリューム番号、そのボリュームが原本であるのか複製であるのかを示すフラグ及び、ローカルノードから各ノードが読み出し可能なのか書き込み可能なのか等の使用状況を示す使用状況情報を含む。ストレージセット管理テーブル509内の情報は、各ノードに備えられた制御装置Cの間で交換される。なお、図7ではボリュームが原本であるのか複製であるのかを示すフラグが「O」である場合ボリュームは原本であり、「C」である場合ボリュームは複製である。なお、後述の逐次格納の場合、格納途中の不完全なボリュームデータがストレージSに格納され得る。この不完全データを示すフラグを、「O」と「C」とは区別できるように、例えば「Q」とすることとしてもよい。

#### 【0076】

例として、図7中のストレージセット番号が「00000001」であるストレージセット構造情報について説明する。このストレージセット構造情報の全体プロパティの読み出しノード数として「3」が格納されているため、データを復元するためには3つのボリュームが必要である。また、同様に、書き込みのノード数として「4」が格納されているため、データは4つのボリュームに分割されている。また、このストレージセット構造情報においてノードA、B、C、E及びG



についてのプロパティが書き込まれているため、これらのノードにポリウムが格納されている。例えば、ノードAのプロパティによると、ノードAにはポリウム番号が「1」であるポリウムが格納され、それは原本であり、読み出し（Read）及び書き込み（Write）が可能な状況であることがわかる。

#### 【0077】

また、ストレージセット構造情報中の使用状況情報は、そのストレージセット管理テーブル509を備えるノードにおいて通常使用されるストレージS、つまりそのノードから見たストレージセットを示す。本実施形態では、ローカルノードから見たストレージセットは、使用状況情報が「RW」つまり、読み出し及び書き込みが可能となっている複数のストレージであるとして仮定する。図7には、例としてノードAに備えられるストレージセット管理テーブル509を示す。この図7によると、ストレージセット番号が「00000001」であるデータについては、ノードAから見たストレージセットは、ノードA、B、及びEに備えられたストレージSである。

#### 【0078】

次に、図8を用いて、アクセス管理テーブル510のデータ構造について説明する。アクセス管理テーブル510は、アクセス単位ごとにアクセス管理情報を格納する。制御装置Cは、アクセス管理情報に基づいて、利用者からの各ノード内のストレージSへのアクセスを制御する。同じストレージセットを共有する利用者からのアクセスは、同じ情報に基づいて制御される。アクセス管理情報は、ストレージセットアクセス番号及び、そのアクセス単位のプロパティを項目として含む。ストレージセットアクセス番号は、ストレージセットへの論理的なアクセス単位（論理ブロック）を示す番号であり、プロパティは、ストレージセットアクセス番号によって示される論理ブロックの状態、例えば、読み出し可能状態であるのか、書き込み可能状態であるのか、ロック状態であるのか、データが完全データであるのか。生成データであるのか等を示す。図8では、読み出し可能状態を「R」で例示し、書き込み可能状態を「W」で例示し、ロック状態を「L」で例示し、原本データを「O」で例示している。なお、ロック状態とは、書き込みが制限される状態をいい、例えばストレージ上のデータが更新される場合等

に、制御装置Cによって設定される（後述）。図8には、例として、ストレージセット番号が「000010001」であるストレージセットについてのアクセス管理テーブルが示されている。また、完全データとは、ストレージから読み出したボリウムから復元されるデータであり、生成データとは、一部のボリウムが足りない状態で冗長化を利用して生成されたボリウムをいう。

#### 【0079】

なお、アクセス管理情報は、さらに、ロックキーを項目として含むこととしてもよい。ロックキーは、ストレージセット番号によって識別されるデータの更新を要求した利用者を識別するための情報であり、更新要求を受信した制御装置Cによって生成される。

#### 【0080】

次に、図9を用いて、ローカルボリウム管理テーブル511のデータ構造について説明する。ローカルボリウム管理テーブル511は、そのテーブルを備える制御装置Cに接続されているストレージS、つまりローカルストレージに格納されているボリウムの利用状況を管理するためのテーブルである。ローカルボリウム管理テーブル511は、各ノードの制御装置Cに個別に存在する。

#### 【0081】

ローカルボリウム管理テーブル511は、ローカルストレージへのアクセス単位ごとにボリウム管理情報を格納する。図9に示すように、ボリウム管理情報は、ローカルストレージへのアクセス単位を示すストレージアクセス番号、そのストレージアクセス番号が示す論理ブロックの状態を示すプロパティ、ストレージセットを識別するストレージセット番号、そのストレージアクセス番号に対応するストレージセットアクセス番号を項目として含む。

#### 【0082】

以下、広域分散ストレージシステムにおける動作について説明する。以下の説明において、データは、2データ+1パリティの冗長構成をとって3つに分割されて、広域分散ストレージシステムに分散格納されると仮定する。しかし、これは説明を分かり易く、且つ、具体的にするためであり、データの冗長構成を限定する趣旨ではない。

## 【 0 0 8 3 】

広域分散システムの利用者は、通常、ネットワーク的に最も近くに位置するノードを介して、広域分散システムにアクセスする。

以下、図 1 0 を用いて、利用者 A がノード A を介して広域分散システムにアクセスする場合のデータの流れについて説明する。図 1 0 において、実線の矢印は利用者が感じるデータの流れを示し、破線の矢印は、実際のデータの流れを示す。利用者 A が、ノード A を介して広域分散システムにアクセスし、データの格納指示を出したと仮定する。

## 【 0 0 8 4 】

利用者 A から見た場合、ノード A に備えられた 1 つの仮想ディスクにデータが格納されたように感じられる。しかし、実際は、ノード A の制御装置 C は、データにパリティを付して複数のボリュームに分割し、広域分散システムを構成する複数のノードに備えられたストレージ S に、ボリュームを分散させて書き込む。図 1 0 の場合、ノード A の制御装置 C は、データを 3 つのボリュームに分割し、それぞれを、ノード A のストレージ S ( A ) 、ノード B のストレージ S ( B ) 及びノード G のストレージ S ( G ) に書き込む。

## 【 0 0 8 5 】

以下、データ書き込みの際の制御装置 C の動作についてより詳しく説明する。

1) 制御装置 C のパケット解析部 3 0 1 は、受信されたパケットを解析し、そのパケットから書き込み指示を示す制御情報とデータを取得する。

## 【 0 0 8 6 】

2) データ分割部 3 0 2 は、データをデータブロック単位に分割する。

3) パリティ計算部 3 0 3 は、パリティを計算し、データブロックに付加する。

## 【 0 0 8 7 】

4) 制御部 5 の経路管理部 5 0 4 は、データブロックを 3 つのボリュームに割り振る事によってデータを 3 分割し、任意の方法で 3 つのノードをそれぞれのボリュームを分散格納するべきストレージセットとして決定する。この説明では、ノード A、ノード B 及びノード G がストレージセットとして決定される。

【 0 0 8 8 】

5) ストレージセット管理部 5 0 5 は、ストレージセットの決定結果に基づいて、ストレージセット構成情報を作成し、ストレージセット管理テーブル 5 0 9 に書き込む。

【 0 0 8 9 】

6) 制御パケット生成部 5 0 2 は、データ書き込み制御を指示する制御パケットを生成する。ネットワーク制御部 5 0 3 は、経路管理部 5 0 4 からの出力に基づいて、パケットの宛先となるノード A、B 及び G を示す経路情報及びそのノードのアドレス等を生成する。

【 0 0 9 0 】

7) データ管理情報付加部 4 0 1 は、3つのボリウムのデータブロックにストレージセット構成情報を付加し、制御経路情報付加部 4 0 2 は、データブロックに、書き込み制御を指示する制御情報及び経路情報を付加する。データ転送部 4 0 3 は、データパケット又は、制御部 5 から出力された制御パケットを転送する。

【 0 0 9 1 】

なお、制御／経路情報に基づいて、データがローカルストレージ、つまり、そのノード A 内のストレージ S (A) に書き込まれるべきデータ（ローカルデータ）であると判定される場合、データ転送部 4 0 3 は、そのデータをパケット解析部 7 に出力する。

【 0 0 9 2 】

8) パケット解析部 7 は、複数のボリウムのうちの1つをローカルストレージ S (A) へ書き込む制御を行う。書き込み後、ローカルボリウム管理部 5 0 6 は、ボリウム管理情報を生成し、ローカルボリウム管理テーブル 5 1 1 に格納する。なお、ボリウム管理情報に含まれる値の内、プロパティ、ストレージセット番号は、パケットから読み出すことにより取得される。

【 0 0 9 3 】

9) 転送されたボリウムは、各転送先のノードのパケット解析部 7 によってそのノードのストレージ S に書き込まれる。格納先のノードのローカルボリウム管

理部 5 0 6 は、上記と同様にしてボリウム管理情報を生成し、ローカルボリウム管理テーブル 5 1 1 に格納する。

【 0 0 9 4 】

一方、利用者が分散格納されたデータを読み出す場合、利用者 A から見た場合、ノード A に備えられた 1 つの仮想ディスクからデータが読み出されたように感じられる。しかし、実際は、ノード A の制御装置 C は、複数のノードから 3 つのボリウムを読み出して、データを復元する。以下、データを読み出す場合の制御装置 C の動作についてより詳しく説明する。

【 0 0 9 5 】

1) 制御装置 C のパケット解析部 3 0 1 は、受信されたパケットを解析し、そのパケットから読み出し指示を示す制御情報を取出す。

2) ストレージセット管理部 5 0 5 は、ストレージセット管理テーブル 5 0 9 から読み出し指示がされたデータのストレージセット構成情報を取得し、これにより、利用者がアクセスしているノードから見てストレージセットとなっているノードのノード名を取得する。この場合は、ノード A、B 及び G である。

【 0 0 9 6 】

3) ノード A のデータ組立部 6 0 2 は、ローカルストレージ S からボリウムを取得する。

4) ノード B 及び G に格納されている残りの 2 つのボリウムを取得するために、制御パケット生成部 5 0 2 は、データ読み出し制御を指示する制御パケットを生成する。ネットワーク制御部 5 0 3 は、パケットの宛先となるノード B 及び G を示す経路情報及びそのノードのアドレス等を生成する。

【 0 0 9 7 】

5) データ管理情報付加部 4 0 1 は、制御パケットにストレージセット構成情報を付加し、制御経路情報付加部 4 0 2 は、データブロックに、書き込み制御を指示する制御情報及び経路情報を付加する。データ転送部 4 0 3 は制御パケットを転送する。

【 0 0 9 8 】

6) ノード B 及びノード G のそれぞれにおいて、転送パケット構築部 4 0 4 は

、制御パケットに基づいてストレージSからボリウムを読み出し読み出されたデータをノードAの制御装置Cに転送するための転送パケットを構築し、データ転送部403に出力する。データ転送部403はパケットをノードAに転送する。

## 【0099】

7) ノードAのパケット解析部7は、受信したパケットから、ノードBのストレージ及びノードGのストレージから読み出された各ボリウムを取得する。

8) パリティ計算部603は、パリティを計算し、データ組立部602は、パリティ及びボリウム番号に基づいて、3つのボリウムから分割される前のデータを組み立てる。パケット構築部601は、データの読み出し指示を出した利用者に、組み立てられたデータを送信するためにパケットを生成する。

## 【0100】

このようにしてデータを複数のボリウムに分割し、ネットワークを介して分散して存在する複数のストレージに各ボリウムを格納させることにより、以下の効果が得られる。

## 【0101】

・1つストレージが盗難にあった場合でも、そのストレージに格納されている1つのボリウムだけでは分割前の元のデータを復元することができないため、データの安全性が高くなる。

## 【0102】

・各ノード宛てのパケットはデータの一部でしかないため、ネットワーク経路においてパケットキャプチャリングを行っても、分割前の元のデータを復元することができない。

## 【0103】

・ストレージが分散配置されるため、ネットワーク的に負荷分散を行う事ができる。このため、従来の技術と同一の速度でバックボーンが構成されている場合は、データアクセスにかかる時間を短縮する事ができる。また、従来の技術と同一のレスポンスを維持する場合は、バックボーンに必要な帯域幅を低減することができる。

## 【0104】

・分散配置と保管が同時に行われるため、バックアップセンタを設けるより、ストレージの使用効率がよい。

上記において、複数のボリウムに分散して格納されたデータを構成する全てのボリウムを複数のノードのストレージSから読み出してデータを復元する際の処理について説明した。しかし、複数のボリウムのうち冗長化された分のボリウムが無くともデータを復元する事ができるため、冗長化された分のボリウムを除いたボリウムを複数のノードのストレージSから読み出すこととしても良い。より具体的には、2データ+1パリティで冗長化されて3つのボリウムに分割されたデータの場合、3つのボリウムのうち2つのボリウムがあればデータを復元できるため、3つのボリウムのうち2つのボリウムをストレージSから読み出して、データを復元することとしてもよい。この場合、ネットワークにかかる負荷をさらに低減する事ができる。

#### 【0105】

ここで、冗長化された分のボリウムを除いたボリウムを複数のノードのストレージSから読み出す場合、読み出されるボリウムの組合せは幾通りも考えられる。以下、このような場合に、最適なボリウムの組合せを決定する方法について説明する。

#### 【0106】

この場合、読み出すべきストレージを決定するために、制御装置C内の経路評価テーブル507に格納される経路評価情報に区間の利用優先度を更に含み、ストレージ評価テーブル508に格納されるストレージ評価情報に経路評価値をさらに含む。

#### 【0107】

以下、図11を用いて利用優先度及び評価値の計算処理の手順について説明する。利用優先度（区間の利用優先度とストレージの利用優先度）及び評価値（経路評価値とストレージ評価値）は、ネットワーク構成が変更されたり、回線が断絶したり、ノードが追加・削除されたりして経路評価テーブル507に格納される情報が変更された場合に、全てのノードについて計算される。

#### 【0108】

図11に示すように、まず、経路管理部504は、利用優先度及び評価値の計算対象として1つのノードを取出し、そのノードがローカルノード（自ノード）であるか否か判定する（S11）。計算対象のノードがローカルノードで無い場合（S11:No）、S12に進み、計算対象のノードがローカルノードである場合（S11:Yes）、S16に進む。

#### 【0109】

S12において、経路管理部504は、計算対象ノードからそのノードに隣接する他のノードまでの各区分について、その区分の帯域幅、コスト、距離を経路評価テーブル507から取得する。さらに、経路管理部504は、区分の利用優先度及びストレージの利用優先度を計算し、その計算結果に基づいて経路評価テーブル507を更新する（S13）。なお、区分の利用優先度及びストレージの利用優先度の計算方法については既に説明した。

#### 【0110】

経路管理部504は、計算対象ノードから他のノードまでの各経路について、経路評価値及びストレージ評価値を計算し、その計算結果に基づいてストレージ評価テーブル508を更新する（S14）。さらに、経路管理部504は、全てのノードについて利用優先度及び評価値を計算したか否か判定する（S15）。全てのノードについて計算を行った場合（S15:Yes）、処理を終了し、そうでない場合、S11にもどる。

#### 【0111】

S16において、経路管理部504は、利用優先度及び評価値を最大値に設定し（S16）、S15に進む。このように、利用優先度及び評価値を最大値に設定することにより、ボリュームの書き込み又は読み出しにおいて、ローカルノードのストレージSは最優先されることになる。

#### 【0112】

図12(a)に、区分の利用優先度を含む、ノードAについての経路評価テーブル507の一例を、図12(b)に、経路評価値を含む、ノードAについてのストレージ評価テーブル508の一例を示す。図12に示すテーブルがノードAについてのテーブルであることは、ノードAが「ローカル」として示されている



ことから分かる。なお、図 1 2 (a) に示す経路評価テーブル 5 0 7 において、区間 C - D についての区間の利用優先度を基準として、他の区間の利用優先度は正規化されている。

#### 【 0 1 1 3 】

以下、図 1 2 を用いて、経路評価値の計算方法について具体的に説明する。図 1 2 (a) に示すように、区間 A - B 及び区間 B - C についての区間の利用優先度は、それぞれ 3 及び 2 である。この場合、経路 A - B - C についての経路評価値は以下のようにして算出される。

#### 【 0 1 1 4 】

経路 A - B - C についての経路評価値

$$\begin{aligned}
 &= \{ (\text{区間 A - B の区間の利用優先度}) + (\text{区間 B - C の区間の利用優先度}) \\
 &\} \div (\text{ホップ数}) \\
 &= (3 + 2) \div 2 \\
 &= 2.5
 \end{aligned}$$

従って、図 1 2 (b) において、経路 A - B - C についての経路評価値として「2.5」が格納されている。

#### 【 0 1 1 5 】

データを復元する際に、経路管理部 5 0 4 は、ストレージセット管理テーブル 5 0 9 から復元すべきデータのボリュームを格納するストレージ S を備えるノードのノード名を取得し、それらのノードのうち、ストレージ評価テーブル 5 0 8 内の経路評価値が大きいノードのストレージ S から優先してボリュームを読み出す事として決定する。読み出しボリュームの読み出しの際は、保管の安全性を考慮する必要は無いため、ディスタンスが考慮されていない経路評価値に基づいてボリュームを読み出すべきストレージを決定することは合理的である。

#### 【 0 1 1 6 】

以下、より具体的に、ノード A、B 及び G のストレージに 3 つのボリュームを分散格納した場合に、経路管理部 5 0 4 がボリュームを読み出すべきストレージを決定する方法について図 1 2 (b) を用いて説明する。ここで、ノード A にアクセスした利用者に復元したデータを送信すると仮定する。

## 【0 1 1 7】

図 1 2 (b) に示すように、ノード A、B 及び G についての経路評価値は、それぞれ「最大」、「3」及び「1 0」である。3 つのボリウムのうち 2 つのボリウムがあればデータを復元する事ができるので、この場合、ノード A に備えられた制御装置 C 内の経路管理部 5 0 4 は、ノード A のストレージ及びノード G のストレージから 1 つずつボリウムを読み出すことを決定する。これにより、ネットワーク上の使用帯域を削減しつつ、良いレスポンスでボリウムをストレージから読み出して、データを復元することが可能となる。

## 【0 1 1 8】

上記広域分散ストレージシステムにデータを分散させて格納する場合に、ノードの数がボリウムの数よりも多いことがよくある。この場合、いずれのノードのストレージ S にボリウムを格納するのか選択することが可能である。以下、最適なストレージセットを決定する方法について説明する。

## 【0 1 1 9】

まず、基本的な考え方について説明する。

ボリウムをネットワーク上に分散されたストレージに格納する場合、帯域幅、コスト及びノード間のディスタンスを考慮することが望ましい。つまり、回線帯域が広く、コストが安いことに加えて、災害からの早期復旧のためにノード間の物理的な距離が離れていることが望ましい。近接したノードの場合、1 つの災害のため同時に障害が発生する事がありうるからである。経路評価テーブル 5 0 7 に格納されるストレージ利用優先度及びストレージ評価テーブル 5 0 8 に格納されるストレージ評価値は、上記の考え方に基づいて、回線帯域が広い程大きな値となり、コストが安い程大きな値となり、ノード間の物理的な距離が離れている程大きな値となるように定義されている。なお、ストレージ利用優先度及びストレージ評価値の計算方法については既に説明した。

## 【0 1 2 0】

以下、図 1 3 を用いて、上記ストレージ評価値に基づいてボリウムを格納するストレージセットを決定する処理について説明する。この処理は利用単位ごとに行われる。なお、以下の説明において、ストレージ評価テーブル 5 0 8 に経路評

価値が項目として含まれていると仮定する。

#### 【0 1 2 1】

利用者から新規のデータの格納指示を受けた場合、ストレージセットを新規に決定することが必要となる。なお、ここでいうストレージセットとは、利用者がアクセスしているノード、つまりローカルノードから見たストレージセットである。ストレージセットとして決定されるべきノードの数は、通常、データを分割する事により得られたボリウムの数と同数である。

#### 【0 1 2 2】

ストレージセットを決定するために、まず、ローカルノードの経路管理部 5 0 4 は、ストレージセット番号を割り当て、ストレージセット構成情報をストレージセット管理テーブル 5 0 9 に格納する (S 2 1)。なお、この時点では、ストレージセット構成情報にはストレージセット番号が割り当てられているだけであり、中身は空である。

#### 【0 1 2 3】

つづいて、経路管理部 5 0 4 は、ストレージ評価テーブル 5 0 8 を参照し、まだストレージセットを構成するノードとして決定されていないノードの中から、最大のストレージ評価値と、その評価値を持つノードのノード名を取得する (S 2 2)。

#### 【0 1 2 4】

経路管理部 5 0 4 は、S 2 2 において同一のストレージ評価値を持つ複数のノードを取得したか否か判定する (S 2 3)。同一のストレージ評価値を持つ複数のノードを取得した場合 (S 2 3 : Y e s)、S 2 4 に進み、そうでない場合 (S 2 3 : N o)、S 3 0 に進む。

#### 【0 1 2 5】

S 2 4 において、経路管理部 5 0 4 は、更に、同一のストレージ評価値を持つノードの数は、不足しているノードの数以上であるか否か判定する。同一のストレージ評価値を持つノードの数が、不足しているノードの数以上である場合 (S 2 4 : Y e s)、S 2 5 に進み、そうでない場合 (S 2 4 : N o)、S 3 1 に進む。なお、不足しているノード数とは、ストレージセットを構成するノードとし

て決定されるべきノードの数から、ストレージセットを構成するノードとして決定されたノードの数を減算した後に残る数である。つまり、不足しているノード数とは、ストレージセットを構成するノードとして決定すべきノードの総数のうち、まだ決定されていないノードの数をいう。

## 【 0 1 2 6 】

S 2 5 において、経路管理部 5 0 4 は、S 2 2 で取得された複数のノードについて、ローカルノードからそのノードに至るまでのホップ数が互いに同じであるか否かを判定する。ホップ数が互いに同じである場合 ( S 2 5 : Y e s ) 、 S 2 6 に進み、そうでない場合 ( S 2 5 : N o ) 、 S 3 2 に進む。

## 【 0 1 2 7 】

S 2 6 において、経路管理部 5 0 4 は、ストレージセット管理テーブル 5 0 9 から S 2 2 で取得された複数のノードの経路評価値を取得し、これらのノードの経路評価値が互いに同じであるか否かを判定する。複数のノードの経路評価値が互いに同じである場合 ( S 2 6 : Y e s ) 、 S 2 7 に進み、そうでない場合 ( S 2 6 : N o ) 、 S 3 3 に進む。

## 【 0 1 2 8 】

S 2 7 において、経路管理部 5 0 4 は、S 2 2 で取得した複数のノードから不足しているノード数と同数のノードを任意に選択し、各ノードのストレージに格納するボリウムを決定する。そして、経路管理部 5 0 4 は、ストレージセット構成情報中の各ノードに対応するフィールドに決定したボリウム番号を書き込む。その際に、経路管理部 5 0 4 は、ボリウムが原本であるか否かを示すフラグ（この場合は原本）及び使用状況情報（書き込み可能及び読み込み可能）も書き込む。これにより、S 2 2 で取得されたノードはストレージセットを構成するノードとして決定され、ノードの状態は利用状態となる。

## 【 0 1 2 9 】

続いて、経路管理部 5 0 4 は、ストレージセットを構成するために必要な数だけノードを決定したか否かを判定する ( S 2 8 ) 。必要な数の分だけノードを決定した場合 ( S 2 8 : Y e s ) 、処理を終了し、そうでない場合 ( S 2 8 : N o ) 、 S 2 2 にもどる。

## 【 0 1 3 0 】

S 3 0 において、経路管理部 5 0 4 は、S 2 2 で取得されたノードを、ストレージセットを構成するノードとして決定する。経路管理部 5 0 4 は、S 2 7 と同様にして、そのノードに備えられたストレージに書き込むボリウムのボリウム番号を決定し、決定結果、ボリウムが原本であるか否かを示すフラグ及び使用状況情報を S 2 1 で作成したストレージセット構成情報に書き込む。その後、S 2 8 に進む。

## 【 0 1 3 1 】

S 3 1 において、経路管理部 5 0 4 は、S 2 2 で取得された複数のノードを、ストレージセットを構成するノードとして決定する。経路管理部 5 0 4 は、S 2 7 と同様にして、S 2 1 で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、S 2 8 に進む。

## 【 0 1 3 2 】

S 3 2 において、経路管理部 5 0 4 は、S 2 2 で取得された複数のノードのうち、ホップ数が少ない方のノードを、ストレージセットを構成するノードとして決定する。経路管理部 5 0 4 は、S 2 7 と同様にして、S 2 1 で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、S 2 8 に進む。

## 【 0 1 3 3 】

S 3 3 において、経路管理部 5 0 4 は、S 2 2 で取得された複数のノードのうち、経路評価値が大きい方のノードを、ストレージセットを構成するノードとして決定する。経路管理部 5 0 4 は、S 2 7 と同様にして、S 2 1 で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、S 2 8 に進む。

## 【 0 1 3 4 】

上記のようにして経路管理部 5 0 4 はストレージセットを構成するノードを決定し、ストレージセット構成情報が作成される。決定結果に基づいて、複数のボリウムは、ストレージセットとして決定されたノードに備えられたストレージに分散格納される。また、このストレージセット構成情報は、各ノードの制御装置

Cに送信され、ストレージセット管理テーブル509に格納される。なお、ストレージ評価テーブル508に経路評価値が含まれていない場合、上記処理において、S26及びS33は行われず。ストレージセットは、ネットワーク構成が変化した場合等に、更新することが可能である。この場合、経路管理部504は、更新されるストレージセットについてのストレージセット構造情報の使用状況情報をクリアした後、S22以降を行う。

## 【0135】

上記のように、制御装置Cは、データを複数のボリュームに分割し、帯域幅やコストだけでなく、ノード間の物理的距離に基づいて選択されたストレージにそれらのボリュームを格納する。これにより、災害により1つのボリュームを格納するストレージが破壊等された場合であっても、他のストレージに格納されたボリュームが無事であれば、データを復元する事が可能となる。従って、十分にノード間の物理的距離があれば、特にデータをバックアップするためのバックアップセンタを備える必要がなくなるという効果が得られる。

## 【0136】

以下、3データ+1パリティで冗長化したデータを複数のストレージに分散格納する場合を例として、ストレージセットの決定方法について具体的に説明する。なお、利用者はノードAにアクセスしていると仮定する。

## 【0137】

この場合、格納されるべきデータは、4つのボリュームに分割される。従って、ノードAに備えられた制御装置C内の経路管理部504は、ストレージ評価テーブル508に格納されたストレージ評価値に基づいて、これらのボリュームを分割格納するべき4つのストレージを決定する。図14(a)に、経路評価テーブル507の一例を、図14(b)に、図14(a)に示す経路評価テーブル507中のデータに基づいて算出されたストレージ評価値を示す。

## 【0138】

図14に即して説明すると、経路管理部504は、最も大きなストレージ評価値を持つノードのストレージから順に4つのストレージ、つまり、ノードA、B、E及びGのストレージを、ボリュームを格納するべきストレージとして決定する。

。経路管理部504は、決定結果に基づいてストレージセット構成情報を作成する。

【0139】

上記のように、冗長化され分割されたデータを構成する各ボリュームは、広域分散ストレージシステムに分散して格納される。さらに、各ボリュームの複製を作成し、ストレージに格納することとしても良いことはいうまでも無い。

【0140】

次に、ストレージセットを決定する際にローカルノードとなったノード以外のノードにアクセスする利用者が、そのストレージセットに格納されるデータを利用する場合について説明する。

【0141】

以下、データの格納を指示し、ストレージセットを決定する際にアクセスした利用者を利用者A、その利用者AがアクセスするノードをノードA、そのストレージセットに格納されるデータを利用する新たな利用者を利用者E、その利用者EがアクセスするノードをノードEと仮定して説明する。

【0142】

利用者Eは、ノードEを介して、広域分散ストレージシステムからデータを取得することができるが、上記のストレージセットは、ノードAから見て利用効率が良いように最適化されている。そこで、新たな利用者Eが利用するノードEから見ても良好な利用効率が得られるように、ボリュームの複製を作成することが可能である。以下、この処理を利用者の追加処理という。

【0143】

図15に、ストレージセットを利用する利用者を追加する処理の手順を示す。以下、図15を用いて利用者の追加処理について説明する。以下の処理は、利用者が追加されるノードの経路管理部504が行う。なお、以下の説明において、ストレージ評価テーブル508に経路評価値が項目として含まれていると仮定する。

【0144】

まず、経路管理部504は、利用者が追加されるストレージセットを特定する

ストレージセット番号を取得する。このストレージセット番号は、例えば追加される利用者がアクセスした際に入力することとしても良い。

#### 【0145】

経路管理部504は、ストレージセット管理テーブル509から、そのストレージセット番号に対応するストレージセット構成情報を取得する(S41)。続いて、経路管理部504は、ストレージセットの決定処理を行う(S42)。この処理は、図13を用いて説明した処理と同様である。

#### 【0146】

経路管理部504は、S41で取得したストレージセット構成情報に基づいて、S42で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されているか否か判定する(S43)。例えば、データが4つのボリウムに分割されている場合、S42でストレージセットとして4つのノードが決定されるが、これらの4つのノードに、既に4つのボリウムが格納されているか否か判定する。

#### 【0147】

S42で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されている場合(S43: Yes)、処理を終了する。この場合、新たに追加される利用者が利用するノードにおいても、良好な利用効率を得られるからである。

#### 【0148】

S42で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されていない場合(S43: No)、経路管理部504は、ストレージセット構造情報に基づいて、S42で決定されたストレージセットを構成するノードのうちでボリウムを格納していないノード(以下、未使用ノード)のノード名及び、不足しているボリウムを格納しているノード(以下、既存ノード)のノード名を取得する(S44)。

#### 【0149】

経路管理部504は、ストレージ評価テーブル508から、未使用ノードおよび既存ノードについてのストレージ評価情報を取得し、各ストレージ評価情報に



含まれるホップ数を比較する (S45)。未使用ノードのホップ数の方が、既存のノードのホップ数よりも小さい場合 (S45: Yes)、S48に進み、そうでない場合 (S45: No)、S46に進む。

## 【0150】

S46において、経路管理部504は、更に、各ストレージ評価情報に含まれる経路評価値を比較する (S46)。未使用ノードの経路評価値に定数 a (1以上) を掛算した値の方が、既存のノードの経路評価値よりも小さい場合 (S46: Yes)、S48に進み、そうでない場合 (S46: No)、S47に進む。

## 【0151】

S47において、経路管理部504は、更に、各ストレージ評価情報に含まれるストレージ評価値を比較する (S46)。未使用ノードのストレージ評価値に定数 b (1以上) を掛算した値の方が、既存のノードのストレージ評価値よりも小さい場合 (S47: Yes)、S48に進み、そうでない場合 (S47: No)、処理を終了する。現在の状態でも、追加される利用者にとって良好な利用効率が得られるからである。

## 【0152】

S48において、経路管理部504は、不足しているボリウムを既存ノードのストレージから複製し、その複製を未使用ノードのストレージに書き込むことと決定する。この結果に基づいて、制御パケット生成部501は、決定結果に基づいて、既存ノード宛てに、ストレージセット番号、ボリウム番号、及び未使用ノードのノード名を含み、制御内容が「ボリウムの複写」である制御パケットを生成し、そのパケットは制御装置Cから送信される。このパケットに基づいて、ボリウムの複製が、未使用ノードに生成される。

## 【0153】

続いて、経路管理部504は、S41で取得されたストレージセット構造情報中の未使用ノードのプロパティに、複写されたボリウムのボリウム番号、ボリウムが複写であることを示すフラグ及び使用状況情報を追記し、処理を終了する。これにより、追加される利用者にとっても良好な利用効率が得られるようになる。

## 【0154】

なお、ストレージ評価テーブル508に経路評価値が含まれていない場合、上記処理において、S46は行われず。また、評価の順を、構成によって変更することとしても良い。

## 【0155】

以下、4つのボリュームにデータを複数のストレージに分散格納する場合を例として、利用者の追加処理について具体的に説明する。説明において、既存の利用者AはノードAにアクセスし、追加される利用者EはノードEにアクセスすると仮定する。

## 【0156】

図16は、利用者の追加処理を説明する図である。図16には、向かって左側に、ノードAから見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているボリュームを示す表が、右側に、ノードEから見たそれらを示す表が記載されている。図16において、左向きの矢印は、「左の表と同じ」ことを示す。

## 【0157】

図16に即して説明すると、図16の左側のノードAから見た表において、最もストレージ評価値が高い4つのノードA、B、E、Gにそれぞれボリュームa、b、c及びdが格納されている。このことから、ストレージセットはノードAから見て利用効率が良いように最適化されていることがわかる。一方、追加される利用者EのノードEから見て最もストレージ評価値が高い4つのノードは、ノードA、B、D及びEである。上記の通り、ノードA、B及びEには、既にボリュームa、b及びcが格納されているが、ノードEには、ボリュームが格納されていない。

## 【0158】

この場合、未使用ノードはノードDとなり、既存ノードはノードGとなり、不足するボリュームはdとなる。

ここで、図16によると、ノードEからノードDまでのホップ数は「1」であり、ノードEからノードGまでのホップ数は「2」である。従って、ノードEの

経路管理部504は、ノードEにポリウムdの複製d'を複写する事により、ノードEから見ても利用効率が良いうようにストレージセットを最適化することに決定する。

#### 【0159】

利用者の追加処理についてさらに説明する。上記の図16の説明において、既存の利用者AがノードAにアクセスするように最適化された広域分散ストレージシステムに、ノードEを利用する利用者Eを追加する処理について説明した。次に、さらに第3の利用者としてノードCを利用する利用者Cを追加する処理について説明する。図17は、第3の利用者の追加処理を説明する図である。図17には、向かって左側に、ノードAから見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているポリウムを示す表が、中心に、ノードEから見たそれらを示す表、右側に、ノードCから見たそれらを示す表が記載されている。図17において、左向きの矢印は、「左の表と同じ」ことを示す。また、この説明においても、格納されるべきデータは、4つのポリウムに分割されると仮定する。

#### 【0160】

第3の利用者を追加する場合も、図17で説明した処理と基本的に同様の処理を行う。つまり、経路管理部504は、ストレージセット管理テーブル509から、利用者が追加されることになるストレージセット番号に対応するストレージセット構成情報を取得し、ストレージセット構成情報に含まれるストレージ評価値が最も高い4つのノードを選択する。図17によると、選択されるノードは、ノードB、C、D及びEである。これらのノードがノードCから見て、ストレージセットを構成するノードとして決定される。

#### 【0161】

続いて、経路管理部504は、決定されたストレージセットを構成するノードに、データを構成するポリウムが全て格納されているか否かを判定する。図17によれば、ノードBにポリウムb、ノードDにポリウムd'（dの複製）、ノードEにポリウムcが書き込まれているが、決定されたストレージセットを構成するいずれのノードにもポリウムaは書き込まれていない事が分かる。また、未使用

ノードはノードCであり、既存ノード、つまりボリウムa又はa'（aの複製）を格納しているノードは、ノードA及びノードFであることが分かる。

## 【0162】

図17の場合、ノードCから見た、未使用ノードと既存ノードのホップ数を比較した結果、既存ノードA及びFのホップ数（それぞれ2及び3）は、未使用ノードCのホップ数（0）よりも大きいため、経路管理部504は、既存ノードからボリウムaを複写して未使用ノードCに格納することとして決定する。

## 【0163】

ここで、既存ノードはノードA及びノードFの2つあるため、複写の方法は以下の2通り考えられる。いずれの方法を採用することとしても良い。

方法1）ホップ数が少なく、且つ評価値が高いノードから不足ボリウムを複写して未使用ノードに格納することとする。図17の場合、ノードAからボリウムaを複写する。

## 【0164】

方法2）既存のノードのうちから2以上のノードを選択し、それらのノードから不足ボリウムを分散させて読み出して複写し、未使用ノードに格納する。図17の場合、ノードA及びノードFからボリウムcを分散させて読み出す。

## 【0165】

次に、分散配置ストレージシステムを構成する各ストレージの状態の確認方法について説明する。各ストレージの状態が正常であるか異常であるかを常時確認することが可能なように、分散配置ストレージシステムを構成することも可能である。この場合、各ノードの制御装置Cはローカルノードのストレージに対して状態を問い合わせ、ストレージはそれに対して正常な状態を示すキープアライブ（keep alive）信号を発信する。ストレージからのキープアライブ信号が途絶えた場合、そのノードの制御装置Cは、ストレージセット管理テーブル509を参照して、そのノードを用いて構成されているストレージセットを特定する。そして、特定されたストレージセットを構成する他のノードに対して、異常が生じたことを通知する。異常を検出した制御装置C及び異常を通知された制御装置C、つまり全ての制御装置Cは、ストレージセット構造情報内の全体プロパティ中の

状態情報を、異常状態を示す値「R (Red)」とする。さらに、全ての制御装置Cは、そのストレージセットへの書き込みを遮断するように設定する。但し、利用者からのデータ読み出し要求に対しては、各制御装置Cは継続して処理を行う。

#### 【0166】

次に、書き込みデータ及び復元データの一時保管について説明する。利用者から格納を指示されたデータは、利用者がアクセスしているローカルノードで冗長化された後に複数のボリュームに分割され、そのローカルノードからストレージセットを構成する各ノードに転送される。この際に、ローカルノードの制御装置Cは、そのデータのストレージセット番号と対応付けて、そのデータを一時記憶領域に保管することとしてもよい。

#### 【0167】

この場合、ローカルノードの制御装置Cは、利用者から、ストレージセット番号とともにデータの読み出し要求を受信した際に、まず、そのストレージセット番号に対応するデータが一時記憶領域に格納されているか否か判定する。

#### 【0168】

データが一時記憶領域に格納されていた場合、制御装置Cは、そのデータを利用者に送信する。そうでない場合、制御装置Cは、ストレージセット番号に対応するストレージセット構造情報をストレージセット管理テーブル509から取得し、そのストレージセット構造情報に基づいて各ノードからデータを復元するために必要なボリュームを取得し、それらのボリュームを用いて復元されたデータを利用者に送信する。このとき、復元されたデータは、制御装置Cの一時記憶領域に格納される。一時記憶領域が一杯になった場合、制御装置Cは、使用頻度の低いデータを削除し、そのデータに使用されていた領域を再利用する。これにより、データ読み出し時のレスポンスを向上させることが可能となる。

#### 【0169】

次に、データをストレージに書き込むタイミングについて説明する。データの格納要求があった場合に、制御装置Cは、すぐにデータを複数のボリュームに分割して複数のノードのストレージに分散格納するのではなく、一旦、データを一時

記憶領域に格納した後に、ストレージに分散格納することとしても良い。

【0170】

より具体的には、制御装置Cは、利用者からデータの格納要求を所定回数受信するのを待ち、その間に格納要求されたデータを一時記憶領域に格納する。データの格納要求を所定回数受信した場合、制御装置Cは、一時記憶領域に格納された各データを複数のボリュームに分割して、ボリュームの1つをローカルノードのストレージに格納させ、他のボリュームそれぞれについては、ストレージセットを構成する他のノードに転送し、各ノードのストレージに格納させる。その後、一時格納領域に書きこまれたデータを消去する。これにより、ローカルノード以外の他のノードにボリュームを転送する回数を低減する事が可能となるため、トラフィックの効率を上げることが可能となる。

【0171】

なお、上記において、制御装置Cは、データの格納要求を所定回数受信するまでデータを一時記憶領域に格納するとして説明したが、データの格納要求を所定回数受信するまで待つ代わりに、所定時間、データを一時記憶領域に格納することとしてもよい。この場合も、上記と同様の効果を得ることが可能である。

【0172】

次に、広域分散ストレージシステムを構成するストレージ上に格納されたボリュームを更新する場合の処理について説明する。上記のように、ストレージ上には、データが複数のボリュームに分割されて格納されている。このようなデータを更新する際に、マルチキャストパケットを利用する事としても良い。以下、この場合の処理について説明する。

【0173】

1) まず、ストレージセットを構成するノードを、ボリュームごとにグループ化する。例えば、図17の右側の表に示すようなストレージセットの場合、ストレージセット管理部505は、ボリュームaのグループとしてノードA、C及びF、ボリュームbのグループとしてノードB、ボリュームcのグループとしてノードE、ボリュームdのグループとしてノードD及びGを定義する。このグループは、不図示のグループテーブルに格納することとしても良い。

## 【0174】

2) 続いて、利用者が各ボリュームを更新する際には、利用者が利用するローカルノードからマルチキャストパケットを用いて、ボリューム a、b、c 及び d の各グループに分けられたノードのストレージに対して書き込みを行う。

## 【0175】

3) ローカルノードのストレージセット管理部 505 は、各ボリュームのグループに分けられたノードから正常に更新が完了した旨の通知を受けると、更新が完了したこととする。

## 【0176】

4) 更新の際に異常が発生した場合、ローカルノードのストレージセット管理部 505 は、異常が発生したノード内のボリュームに対して再度の更新処理を行う。

## 【0177】

さらに、ストレージ上に分割して格納されたデータが利用者によって更新される場合、その更新処理の間、他の利用者がそのデータを構成するボリュームの原本或いは複製を更新することを禁止することとしても良い。これにより、更新されるデータについて原本と複製の内容の統一を図ることが可能となる。以下、この場合の処理について説明する。

## 【0178】

行われる処理の手順の概要は以下のとおりである。

1) まず、利用者はノードにアクセスし、更新したいデータのストレージセット番号を指定し、更新要求をそのノードに送信する。以下、この利用者がアクセスするノードをローカルノードという。ローカルノードの制御装置 C 内のストレージセット管理部 505 は、その利用者に対してロックキーを発行する。ロックキーは、データの更新を要求した利用者を識別するために使用され、広域分散ストレージシステムの利用者及びセッションごとにユニークなものである。図 22 に、このロックキーの機能を追加したアクセス管理テーブル 510 及びローカルボリューム管理テーブル 511 の一例を示す。

## 【0179】

2) ローカルノードの制御装置C内のストレージセット管理部505は、そのストレージセット番号によって特定されるボリウムを他の利用者が更新する事を禁止し(以下、ロックという)、さらに、そのストレージセットを構成する他のノードに、対応するボリウムをロックするよう依頼する。

【0180】

3) 依頼を受信した各ノードの制御装置C内のストレージセット管理部505は、各ボリウムをロックする。

4) ローカルノードの制御装置C内のストレージセット管理部505は、各ボリウムがロックされていることを確認する。

【0181】

5) ローカルノードの制御装置Cは、更新すべきデータを冗長化して複数のボリウムに分割し、各ボリウムを、それぞれを格納すべきノードに送信する。

6) ボリウムの送信及び更新が終了すると、各ノードの制御装置C内のストレージセット管理部505は、ロックを解除する。

【0182】

以下、上記の手順2)から6)を順に詳しく説明する。まず、手順2)について、図18を用いて詳しく説明する。図18に示す手順は、利用者からストレージセット番号の指定と、更新要求を受信したローカルノードの制御装置C内のストレージセット管理部505によって行われる。

【0183】

まず、ストレージセット管理部505は、利用者から受信したストレージセット番号に対応するアクセス管理情報を取得し、そのアクセス管理情報に基づいて、更新要求が出されたアクセス単位(論理ブロック)の状態が、「ロック状態」であるのか否か判定する(S51)。そのアクセス単位の状態が「ロック状態」と判定された場合(S51:Yes)。S52に進み、そのアクセス単位の状態が「ロック状態」以外の状態である場合、S57に進む。S57において、ストレージセット管理部505は、ロックに失敗した旨を利用者に通知し、処理を終了する(終了パターン2:異常終了)。異常終了の場合、更新を行う事はできない。



## 【 0 1 8 4 】

S 5 2 において、ストレージセット管理部 5 0 5 は、ロックキーを生成する。

続いて、ストレージセット管理部 5 0 5 は、ストレージセット番号に対応するストレージセットについてのストレージセット構造情報をストレージセット管理テーブル 5 0 9 から取得し、そのストレージセットを構成するノードに、そのストレージセットをロックする依頼を通知する (S 5 3)。なお、ロック依頼には、ストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。なお、ストレージセットを構成するノードにローカルノードが含まれる場合もあることをいうまでもない。

## 【 0 1 8 5 】

通知を受けた各ノードは、ロック処理を行う (S 5 4)。この処理については後述する。

さらに、ローカルノードのストレージセット管理部 5 0 5 は、S 5 3 で通知した各ノード全てからロックが完了した旨の通知を待つ (S 5 5)。全ノードからロックが完了した旨の通知を受信した場合 (S 5 5 : Y e s)、S 5 8 に進む。そうでない場合 (S 5 5 : N o)、S 5 6 に進む。

## 【 0 1 8 6 】

S 5 6 において、ストレージセット管理部 5 0 5 は、利用者に対してロックに失敗した事を通知し、処理を終了する (終了パターン 2 : 異常終了)。

S 5 8 において、ストレージセット管理部 5 0 5 は、更新要求が出されたアクセス単位についてのアクセス管理情報にふくまれるプロパティを「ロック状態」に更新し、さらに、S 5 2 で生成されたロックキーを追加するように、アクセス管理テーブル 5 1 0 を更新する (S 5 8)。さらに、ストレージセット管理部 5 0 5 は、その利用者に対してロックキーを発行して (S 5 9)、処理を終了する (終了パターン 1 : 正常終了)。

## 【 0 1 8 7 】

続いて、手順 3) について、図 1 9 を用いて説明する。この手順 3) は、図 1 8 の S 5 4 において、ロックの依頼を受信した各ノードの制御装置 C によって行われる処理に相当する。

## 【0188】

まず、ローカルボリューム管理部506は、ロック依頼と共に受信したストレージセット番号及びストレージセットアクセス番号に対応するボリューム管理情報をローカルボリューム管理テーブル511から取得し、そのボリューム管理情報に基づいて、更新要求が出されたアクセス単位（論理ブロック）の状態が、「ロック状態」であるのか否か判定する（S61）。そのアクセス単位の状態が「ロック状態」であると判定された場合（S61：Yes）。S62に進み、そのアクセス単位の状態が「ロック状態」以外の状態である場合、S66に進む。S66において、ローカルボリューム管理部506は、ロックに失敗した旨を利用者に通知し、処理を終了する（終了パターン2：異常終了）。異常終了の場合、更新を行う事はできない。

## 【0189】

S62において、ローカルボリューム管理部506は、更新要求が出されたアクセス単位についてのボリューム管理情報にふくまれるプロパティに「ロック状態」を示すフラグを追加し、さらにロックキーを追加する。続いて、ストレージセット管理部505は、ロックが完了した旨を、ロック依頼を通知してきたノードの制御装置Cに通知する（S63）。

## 【0190】

さらに、ストレージセット管理部505は、アクセス管理テーブル510に、ロック対象となっている論理ブロックについてのアクセス管理情報が格納されているか否か判定する（S64）。アクセス管理情報が格納されている場合（S65：Yes）、ストレージセット管理部505は、そのアクセス管理情報のプロパティを「ロック状態」に更新する。アクセス管理情報がアクセス管理テーブル510に格納されていない場合（S65：No）、処理を終了する（終了パターン1：正常終了）。

## 【0191】

続いて、手順4）から6）までについて図20及び21を用いて説明する。まず、更新要求を出した利用者は、発行されたロックキーとともに、更新したいデータの内容をローカルノードに送信する（不図示）。まず、図20について説明

する。図 2 0 に示す処理は、ローカルノードの制御装置 C によって行われる。

【 0 1 9 2 】

ローカルノードのストレージセット管理部 5 0 5 は、更新対象となる論理ブロック（ロック対象の論理ブロックでもある）についてのアクセス管理情報をアクセス管理テーブル 5 1 0 から取得する。（S 7 1）。続いて、ストレージ管理部 5 0 5 は、S 7 1 で取得したアクセス管理情報に含まれるプロパティに基づいて、更新対象となるブロックはロックされているか否か判定する（S 7 2）。更新対象となるブロックがロックされている場合（S 7 2 : Y e s）、S 7 3 に進み、更新対象となるブロックがロックされていない場合（S 7 2 : N o）、S 7 8 に進む。

【 0 1 9 3 】

S 7 3 において、ストレージセット管理部 5 0 5 は、利用者から受信したロックキーが、アクセス管理情報に含まれるロックキーと一致するか否か判定する。2 つのロックキーが位置しない場合（S 7 3 : N o）、ストレージセット管理部 5 0 5 は、利用者に対し、書き込み（更新）に失敗した旨を通知し（S 7 9）、処理を終了する（終了パターン 2 : 異常終了）。

【 0 1 9 4 】

2 つのロックキーが一致した場合（S 7 3 : Y e s）、データ変換部 3 は、利用者から取得したデータを冗長化し、複数のボリウムに分割する。さらに、ストレージセット制御部 5 0 1、ストレージセット管理テーブル 5 0 9 から、アクセス対象となっているストレージセットについてのストレージセット構造情報を取得する。経路管理部 5 0 4 は、ストレージセット管理テーブル 5 0 9 からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのストレージセットを構成する各ノードに、書き込み依頼と共にボリウムを送信するように制御情報及び経路情報を生成し、パケット生成部 4 は、制御情報及び経路情報を付加したパケットを各ノードに送信する。なお、書き込み依頼にはストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。これを受けて、各ノードでは、データをストレージ S に書き込む処理が行われる（S 7 4）。この処理については図 2 1 を用いて

後述する。

#### 【 0 1 9 5 】

続いて、ストレージセット管理部 5 0 5 は、書き込み依頼を送信した全てのノードから書き込み完了の通知を受信するまで待つ（S 7 5）。全てのノードから書き込み完了の通知を受信しなかった場合（S 7 5 : N o）、ローカルノードの制御装置 C は、そのストレージセットを構成する各ノードに、書き込み依頼と共にボリウムを再び送信する。これを受けて、各ノードでは、データをストレージ S に書き込む処理が再び行われる（S 8 0）。この処理は、S 7 4 と同じである。

#### 【 0 1 9 6 】

全てのノードから書き込み完了の通知を受信した場合（S 7 5 : Y e s）、制御装置 C は、書き込みの要求を出した利用者へ書き込み完了の通知を送信する（S 7 6）。ストレージセット管理部 5 0 5 は、更新対象となる論理ブロックについてのアクセス管理情報をアクセス管理テーブル 5 1 0 から取得し、アクセス管理情報に含まれるプロパティから「ロック状態」を示すフラグを削除し（S 7 7）、処理を終了する（終了パターン 1 : 正常終了）。

#### 【 0 1 9 7 】

S 7 8 において、制御装置 C はロック処理を行う。ロック処理について、上記において図 1 8 及び 1 9 を用いて既に説明したため、ここでは説明をしない。S 7 8 のロック処理が正常終了した場合（終了パターン 1）、S 7 4 に進み、S 7 8 のロック処理が異常終了した場合（終了パターン 2）、S 7 9 に進む。

#### 【 0 1 9 8 】

次に、図 2 1 について説明する。図 2 1 の処理は、上記の図 2 0 の S 7 4 の処理に相当し、書き込み依頼を受信した各ノードの制御装置 C によって行われる。

まず、ローカルボリウム管理部 5 0 6 は、ローカルボリウム管理テーブル 5 1 1 から書き込み依頼と共に受信したストレージセット番号及びストレージセットアクセス番号に対応するボリウム管理情報をローカルボリウム管理テーブル 5 1 1 から取得する（S 8 1）。続いて、ローカルボリウム管理部 5 0 6 は、そのボリウム管理情報に含まれるロックキーが、書き込み依頼に含まれるロックキーと

一致するか否か判定する（S 8 2）。2つのロックキーが一致する場合（S 8 2 : Y e s）、S 8 3に進み、一致しない場合（S 8 2 : N o）、ローカルボリウム管理部 5 0 6 は、書き込みに失敗した旨を、書き込み依頼を送信してきたノードの制御装置 C に送信し、処理を終了する（終了パターン 2 : 異常終了）。

## 【 0 1 9 9 】

S 8 3において、パケット解析部 7 は、書き込み依頼と共に受信したパケットからボリウムを取出して、ストレージ I F 8 を介してそのボリウムをストレージ S に書き込む。続いて、制御パケット生成部 5 0 2 は、書き込みが完了した旨の通知を、書き込み依頼を送信してきたノードの制御装置 C に送信する（S 8 4）。続いて、ローカルボリウム管理部 5 0 6 は、S 8 1 で取得したボリウム管理情報に含まれるプロパティから、「ロック状態」を示すフラグ及びロックキーを削除し（S 8 5）、処理を終了する。

## 【 0 2 0 0 】

次に、上記のロック処理において、更新の際にマルチキャストパケットを用いる場合について説明する。この場合の処理の手順の概要は以下のとおりである。なお、上記のマルチキャストを用いる更新処理の場合と同様に、処理の前にストレージセットを構成するノードを、ボリウムごとにグループ化することが必要である。各ノードの制御装置 C には、各ボリウムのグループを構成するノードとそのグループの代表ノードを示す情報を含む不図示のノードグループテーブルが備えられている。

## 【 0 2 0 1 】

1) ストレージセット中で、各ボリウムを格納するノードを代表する代表ノード（例えば、原本を格納するノード）に対して、データ更新の前にデータアクセス単位でロック処理を行う。

## 【 0 2 0 2 】

2) 原本は同じグループに属するノードに対して、上記更新要求を送信した利用者を確認できるようにして（ロックキーを使用）ロック処理を行う。なお、1) 及び 2) の手順は、上記のロック処理と同様である。

## 【 0 2 0 3 】

3) 更新要求を送信した利用者は、マルチキャストパケットを用いて各ボリウムの更新内容を送信する。

4) 更新内容を含むパケットを受信した各ノードは、ロックキーを用いて更新要求を送信した利用者と、そのパケットを送信した利用者とが同一であることを確認し、確認できた場合、ボリウムの更新を行う。

#### 【 0 2 0 4 】

5) 上記4)における更新の際、ボリウムのグループの代表ノードは、更新完了を示すパケットを利用者がアクセスしているノードに送信する。代表ノード以外のノードは、そのグループの代表ノードに更新完了を示すパケットを送信する。

#### 【 0 2 0 5 】

6) 代表ノードは、自ノードに格納されたボリウムに対するロックを解除する。また、代表ノードは、自グループに属する他のノードから更新完了を示すパケットを受信した場合、そのノードに格納されたボリウムに対するロックを解除する。

#### 【 0 2 0 6 】

7) 代表ノードは、自グループに属する他のノードのうちで、更新完了を示すパケットを送信してこないノードがある場合、そのノードに対して更新処理を実行する。

#### 【 0 2 0 7 】

以下、図 2 3 から図 2 5 を用いて、上記 3) からの手順についてより詳しく説明する。なお、図 2 3 から図 2 5 に示す処理は、図 2 0 及び図 2 1 に示す処理と同様の手順を含むため、図 2 3 から図 2 5 において図 2 0 及び図 2 1 と同様の手順には同じ番号を付し、説明を省略する。まず、図 2 3 を用いて、更新要求を出した利用者がアクセスするノードで行われる処理について説明する。以下において、利用者がアクセスするノードをローカルノードという。

#### 【 0 2 0 8 】

図 2 3 では、図 2 0 の S 7 4 から S 7 7 及び S 8 0 の代わりに、S 9 1 から S 9 5 を行う点が図 2 0 に示す処理と異なる。以下、S 9 1 から S 9 5 について説

明する。なお、S 9 1 から S 9 5 は、上記 3) から 7) に示す手順のうち、ローカルノードの制御装置 C によって行われる手順を示す。

【 0 2 0 9 】

S 7 1 から S 7 3 の後、経路管理部 5 0 4 は、ストレージセット管理テーブル 5 0 9 からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのストレージセットを構成する各ノードに、書き込み依頼と共にそのノードに書き込むべきボリュームを送信するように制御情報及び経路情報を生成し、パケット生成部 4 は、制御情報及び経路情報を付加したパケットを各ノードにマルチキャストで送信する。なお、書き込み依頼にはストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。これを受けて、各ノードでは、ボリュームをストレージ S に書き込む処理が行われる (S 9 1) 。この処理については図 2 4 及び図 2 5 を用いて後述する。

【 0 2 1 0 】

続いて、ストレージセット管理部 5 0 5 は、書き込み依頼を送信したノードのうち、代表ノードから書き込み完了の通知を受信するまで待つ (S 9 2) 。ここで、ストレージセット管理部 5 0 5 は、不図示のノードグループテーブルに基づいて、全ての代表ノードから書き込み完了の通知を受信したか否か判定する。なお、図 2 3 から図 2 5 において原本を格納するノードを代表ノードと仮定している。

【 0 2 1 1 】

全ての代表ノードから書き込み完了の通知を受信しなかった場合 (S 9 2 : N o) 、ローカルノードの制御装置 C は、その代表ノードに、書き込み依頼と共に書き込むべきボリュームを再び送信する。これを受けて、各ノードでは、データをストレージ S に書き込み処理が再び行われる (S 8 0) 。この処理は、S 9 1 と同じである。

【 0 2 1 2 】

全てのノードから書き込み完了の通知を受信した場合 (S 9 2 : Y e s) 、ローカルノードのストレージセット管理部 5 0 5 は、書き込みが完了した旨を利用

者に通知する（S93）。続いて、ストレージセット管理部505は、更新対象となる論理ブロックについてのアクセス管理情報をアクセス管理テーブル510から取得し、アクセス管理情報に含まれるプロパティから「ロック状態」を示すフラグを削除し、処理を終了する（終了パターン1：正常終了）。

#### 【0213】

また、ローカルノードの制御装置Cは、S78及びS79も行う。S78及びS79の処理については図20と同様であるため説明を省略する。

次に、図24を用いて、図23のS91において行われる処理について説明する。図24の処理は、各ボリウムのグループにおける代表ノードによって行われる。

#### 【0214】

図24では、図21のS85の後に、更に、S101及びS102を行う点が図21に示す処理と異なる。以下、S101及びS102について説明する。

S81からS85の後、代表ノードのストレージセット管理部505は、その代表ノードが属するグループ内の他のノードの全てから書き込み完了の通知を受信するまで待つ（S101）。この代表ノードはノードグループテーブル（不図示）に基づいて、S101の判定を行う。全てのノードから書き込み完了の通知を受信しなかった場合（S101：No）、代表ノードの制御装置Cは、書き込み完了の通知を送信してこなかったノードに対するボリウムの書き込みを、そのノードに代わって実行し（S102）、S101に戻る。

#### 【0215】

全てのノードから書き込み完了の通知を受信した場合（S101：Yes）、処理を正常終了する（終了パターン1）。

次に、図25を用いて、図23のS91において行われる処理について説明する。図25の処理は、ストレージセットを構成するノードのうちで、代表ノード以外のノードによって行われる。

#### 【0216】

図25では、図21のS83の後に、S84の代わりに、S111を行う点が図21に示す処理と異なる。以下、S111について説明する。



S 8 1 から S 8 3 の後、代表ノード以外のノードのストレージセット管理部 5 0 5 は、そのノードが属するグループの代表ノードへ書き込み完了の通知送信する (S 1 1 1)。

#### 【 0 2 1 7 】

次に、新規のボリウムの作成手順について説明する。新規のボリウムは、たとえば、ボリウムの複製を作成する際や障害からの復旧の際等において作成される。

#### 【 0 2 1 8 】

以下、ボリウムの複製を作成する際を例にとって、新規のボリウムの作成手順について説明する。図 2 6 は、向かって左から順に、ノード A にアクセスする利用者 A、ノード B にアクセスする利用者 B 及びノード C にアクセスする利用者が、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリウムを示す表である。以下、図 2 6 を用いて、複製の作成方法の決定について説明する。なお、この説明において、データは 4 つのボリウムに分割されていると仮定する。

#### 【 0 2 1 9 】

図 2 6 によると、利用者 C が使用するべきストレージセットは、ノード B、C、D 及び E であり、利用者 C を広域分散ストレージシステムに追加する際に、ノード C にボリウム a の複製が作成される。このボリウム a の複製は、以下の 2 通りの方法で作成することができる。

#### 【 0 2 2 0 】

- 1) ノード A 又は F にのいずれかに格納されたボリウム a から作成する。
- 2) ボリウム a の複製を他のボリウム b、c 及び d から冗長を用いて再現する。

#### 【 0 2 2 1 】

複製が作成されるボリウムを格納するノードと、その他のボリウムを格納するノードのノード C から見たストレージ評価値を比較し、複製が作成されるボリウムを格納するノードの最高ストレージ評価値を、他の各ボリウムを格納するノード

ドの最高ストレージ評価値のそれぞれが上回っていた場合、上記のうち 2) の方法を採用し、そうでない場合、1) の方法を採用する。

#### 【 0 2 2 2 】

例えば、図 2 6 の場合、上位 4 つのストレージ評価値は以下のようになる。

ボリウム a を格納するノードの最高ストレージ評価値：1 1 . 3 (ノード A)

ボリウム b を格納するノードの最高ストレージ評価値：1 7 . 0 (ノード B)

ボリウム c を格納するノードの最高ストレージ評価値：1 4 . 8 (ノード E)

ボリウム d を格納するノードの最高ストレージ評価値：2 1 . 8 (ノード D)

ボリウム a を格納するノード A のストレージ評価値は、他のボリウムを格納するノードのストレージ評価値のいずれよりも低い。従って、この場合、ノード C の経路管理部 5 0 4 は、ノード C 作成されるボリウム a の複製を、それぞれノード B、D 及び E に格納されるボリウム b、d 及び c から冗長を利用して再現することと決定し、この決定に基づいて、各ノードからボリウム b、c 及び d はノード C に転送され、ノード C のパケット解析部 7 は、受信したボリウム b、c、d からボリウム a を再現して、ストレージ I F 8 を介してストレージ S にボリウム a を書き込む。

#### 【 0 2 2 3 】

次に、広域分散ストレージシステムを構成するノードの一部に障害が発生した場合の処理について説明する。まず、ノードに障害が発生した場合に、残っているノードから見て最適なストレージセットを再設定する処理について説明する。説明において、データは 4 つのボリウムに分割されていると仮定する。また、障害発生前後の各ノードの状態を図 2 7 (a) 及び (b) に示すような状態として仮定する。なお、図 2 7 (a) 及び (b) において、向かって左側に、ノード A から見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているボリウムを示す表が、中心に、ノード E から見たそれらを示す表、右側に、ノード C から見たそれらを示す表が記載されている。図 2 7 (a) 及び (b) において、左向きの矢印は、「左の表と同じ」ことを示す。

#### 【 0 2 2 4 】

まず、図 2 7 (a) に示すような状態において、ノード A の利用者 A は、ノー

ドA、B、E及びGからそれぞれボリウムa、b、c及びdを取得する。ノードEの利用者Eは、ノードA、B、D及びEからそれぞれボリウムa、b、d'及びcを取得する。ノードCの利用者Cは、ノードB、C、D及びEからそれぞれボリウムb、a'、d'及びcを取得する。なお、「'」は、そのボリウムが複製である事を示す。

## 【0225】

このような状態で、ノードAのストレージに障害が発生した場合、利用者A及び利用者Eは、ノードAからボリウムaを取得する事ができなくなる。この場合、ボリウムaは、ノードC及Fにも存在するため、ノードA及びノードEのストレージセット管理部は、ノードAの代わりにノードC又はFのいずれかからボリウムaを取得する事に決定する。

## 【0226】

どのノードからボリウムaを取得するのか決定する方法は、基本的に、ストレージセットの決定処理と同様に、以下のように決定する。

- 1) ストレージ評価値が高い方を採用する。

## 【0227】

- 2) ストレージ評価値が同じ場合にホップ数が少ない方を採用する。

図27(b)において、ノードAから見て、ノードCとノードFのストレージ評価値は、それぞれ、「10.8」及び「8.3」である。従って、ノードAのストレージセット管理部505は、ノードAのストレージの代わりにノードCのストレージからボリウムaを取得する事として決定する。同様に、ノードEから見て、ノードCとノードFのストレージ評価値はそれぞれ、「16.3」及び「13.0」である。従って、ノードEのストレージセット管理部505は、ノードAのストレージの代わりにノードCのストレージからボリウムaを取得する事として決定する。

## 【0228】

さらに、障害が生じたノードAのストレージに格納されていたボリウムaは原本であったが、今回の障害の結果、ボリウムaの原本が存在しない事となってしまうため、ノードCとノードFのいずれかを代表ノードとして扱う事と決定する

。なお、図 2 7 (b) においては、各ノードからのノード C 及びノード F に対するストレージ評価値のうち、より平均値が大きいノード C を代表ノードとしている。

#### 【 0 2 2 9 】

さらに、障害からの復旧の際の、新規ボリュームの作成手順について説明する。図 2 8 は、図 2 7 と同様に、向かって左から順に、ノード A にアクセスする利用者 A、ノード E にアクセスする利用者 E 及びノード C にアクセスする利用者 C が、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリュームを示す表である。以下、図 2 8 を用いてノード A のストレージが障害から普及した場合、ノード A のストレージにボリュームを格納する際に、どのボリュームの複製をどうやって作成するのか、その決定手順について説明する。なお、この説明においても、データは 4 つのボリュームに分割されていると仮定する。

#### 【 0 2 3 0 】

1) まず、ストレージセット管理部 5 0 5 は、復旧するノードのストレージ評価値は、復旧前に使用されていたノードのストレージ評価値よりも高いか否か判定する。復旧するノードのストレージ評価値の方が復旧前に使用されていたノードのストレージ評価値よりも高い場合、ストレージセット管理部 5 0 5 は、復旧するノードのストレージ S にボリュームの複製を書き込むことと決定する。複製されるボリュームは、復旧前に使用されていたノードのうち、最低のストレージ評価値を持つノードのストレージに格納されていたボリュームとする。

#### 【 0 2 3 1 】

例えば、図 2 8 に示すように、ノード A では、ノード A のストレージ S の復旧前に、ノード B、C、E 及び G のストレージが使用されていたが、ノード A のストレージ評価値の方がこれらより高い。また、ノード B、C、E 及び G のうち最低のストレージ評価値を持つノードはノード C であり、そのノード C のストレージ S に格納されているボリュームはボリューム a である。同様に、ノード E では、ノード A のストレージ S の復旧前に、ノード B、C、D 及び E のストレージが使用されていたが、このうちの最低のストレージ評価値を持つノード C よりもノード

Aのストレージ評価値の方が高い。ノードCでは、ノードAの復旧前後において、追加されるボリュームを格納するノードのストレージ評価値は、使用するノードの変更を要するほど高くないため、何もしない。

#### 【0232】

2) 複製の作成は、最もストレージ評価値の高いノードで行われる。通常、ボリュームの複製が書き込まれるストレージに直接接続されているノードで行われる。図28の場合、ノードA（つまり、複製が作成されるストレージに接続されているノード）の制御装置Cは、ノードAのストレージSにボリュームaの複製を書き込む。

#### 【0233】

3) 複製を作成する際、ノードAのストレージセット管理部505は、複製される元となるボリュームを格納するノードのストレージ評価値と、その他のボリュームを格納するノードのストレージ評価値とを比較し、比較結果に基づいて、ストレージに格納されたボリュームを元にして複製を作成するか、それとも他のボリュームから冗長を利用してボリュームを再現するか、決定する。この決定方法は、上記の複製の作成方法と同様であるため、詳しい説明は省略する。

#### 【0234】

図28の場合、ボリュームaを格納するノードのうち最大のストレージ評価値を持つノードはノードCであり、その値は10.8であり、この値は、他のボリュームを格納するノードのストレージ評価値よりも低い、従って、ストレージセット管理部505は、他のボリュームb、c及びdから冗長を利用してボリュームaを再現する事と決定する。

#### 【0235】

次に、ストレージセットの管理について説明する。ノードの追加や削除を繰り返し、利用されないボリューム等が存在するようになる。このような場合、利用されないボリュームを削除することにより、ストレージの利用効率を向上させるようにストレージセットを管理ことが可能である。以下、ストレージセットの管理について説明する。

#### 【0236】

以下において、データは4つのポリウムに分割され、ノードの追加や削除の結果、図29に示すようなポリウム構成になっていると仮定する。なお、図29（a）及び（b）は、それぞれ向かって左から順に、ノードAにアクセスする利用者A、ノードEにアクセスする利用者E及びノードCにアクセスする利用者Cが、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているポリウムを示す表であり、これは、図28と同様である。ここで、図29（a）は、使用されていないポリウムの削除前の状態を示し、図29（b）は、使用されていないポリウムの削除後の状態を示す。

## 【0237】

1）一定時間経過ごと、或いは、各ノードのストレージSに格納されるポリウムの状態が変更されるごとに、各ノードにおいて使用されるストレージセットを示す情報を各ノードで交換する。あるいは、任意の1つのノードに、その情報を集める。

## 【0238】

2）交換された情報に基づいて、どのノードによっても使用されていないポリウムが合った場合、1つのノードのストレージセット管理部505は、そのノードのポリウムを削除する事と決定し、そのノードに対してそのポリウムを削除するよう指示する制御パケットを送信する。

## 【0239】

例えば、図29において、ノードAにおいてストレージセットとして使用されているノードはノードA、B、E及びGであり、ノードEにおいてストレージセットとして使用されているノードはノードA、B、D及びEであり、ノードCにおいてストレージセットとして使用されているノードは、ノードB、C、D及びEである。（使用されているノードは最もストレージ評価値が高い4ノードである）従って、ノードFは、どのノードの利用者によっても使用されていないことがわかる。従って、ノードFのストレージSに格納されるポリウムa'は削除される（図29（b）参照）。

## 【0240】

次に、データの複製の再生又はデータの再生を逐次に行う処理について説明する。データの複製又は再生は、利用者が追加された場合や新たなノードが追加された場合等において行われるが、緊急に行う必要が無いことも多い。この場合、データの複製又は再生は、ネットワークの空き時間等を利用して逐次に行われることとしてもよい。これにより、トラフィックの有効利用を可能とする。以下、この場合の処理について説明する。以下の処理は、利用者からデータの読み込み要求又は書き込み要求を受信したノード（以下、ローカルノード）によって行われる。

#### 【 0 2 4 1 】

図 3 0 に、逐次にデータの複製の作成又は再生を行う場合に、処理についてのフローチャートを示す。図 3 0 に示すように、まず、利用者は、読み出したいデータのストレージセット番号を指定して読み込み要求又は書き込み要求をアクセス先となっているノードに送信する。ローカルノードのストレージセット管理部 5 0 5 は、ストレージセット管理テーブル 5 0 9 からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのローカルノードが使用するストレージセットに、データを構成する全てのポリウムが格納されているか否か判定する（S 1 2 1）。

#### 【 0 2 4 2 】

ローカルノードが使用するストレージセットに、データを構成する全てのポリウムが格納されている場合（S 1 2 1 : Y e s）、データの読み込み・書き込み処理が行われる（S 1 2 6）。この読み込み・書き込み処理については、既に説明した。

#### 【 0 2 4 3 】

ローカルノードが使用するストレージセットに、データを構成する全てのポリウムが格納されていない場合（S 1 2 1 : N o）、ストレージセット管理部 5 0 5 は、読み出し要求を受信した場合は、各ノードから取得したポリウムから冗長化によって要求されたデータを生成する。書き込み要求を受信した場合は、ストレージセット管理部 5 0 5 は、受信したデータを冗長化し、複数のポリウムに分割して各ノードに書き込む。その際に、ストレージセット管理部 5 0 5 は、読み

出し又は書き込みアクセス管理テーブル 5 1 0 から該当するストレージセット番号を持つアクセス管理情報を取得し、そのアクセス管理情報に含まれるそのボリュームへのアクセスに関するプロパティに、アクセスされたデータが、ストレージ S から読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付す (S 1 2 2)。

#### 【 0 2 4 4 】

続いて、利用者から読み出し要求を受信した場合、ストレージセット管理部 5 0 5 は、S 1 2 1 で取得したストレージセット構造情報に基づいて、ストレージセットを構成するノードのうちでボリュームを格納していないノード（不完全ノード）を特定し、そのノードに格納するためのボリュームを、ストレージセットを構成する他のノードから読み出されたボリュームから生成する (S 1 2 3)。

#### 【 0 2 4 5 】

ローカルストレージの制御装置 C は、生成されたボリュームを書き込むように指示して、不完全ノードのストレージ S にそのボリュームを転送する。これを受けて、不完全ノードでは、ボリュームの書き込み処理が行われる (S 1 2 4)。この書き込み処理は、逐次にネットワークの空き時間を利用して行われる。なお、ストレージセット管理部 5 0 5 は、このボリュームが完全データでないことを示すフラグを S 1 2 1 で取得したストレージセット構造情報に含まれる不完全ノードについてのプロパティに付す。

#### 【 0 2 4 6 】

逐次書き込みにおいて、その書き込みによるアクセスを示すアクセス管理情報中のプロパティに、アクセスされたデータが、ストレージ S から読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付される。逐次書き込みが完了すると、アクセス管理情報中のプロパティには、生成データであることを示すフラグは付されないことになる。なお、2 度目以降の逐次書き込みにおいて、ローカルノードの制御装置 C は、ストレージセット構造情報及びアクセス管理情報内の生成データを示すフラグに基づいて、不完全ノードと、それに格納すべきボリュームを特定することができる。

#### 【 0 2 4 7 】



書き込み処理が終了すると、ストレージセット管理部 5 0 5 は、アクセス管理テーブル 5 1 0 を参照し、ストレージセット番号に対応するアクセス管理情報に、生成データであることを示すフラグが付されているか否か判定する。生成データであることを示すフラグが付されたアクセス管理情報が無ければ、ローカルノードが使用するストレージセットに、データを構成する全てのボリュームが格納されていることとなる (S 1 2 5 : Y e s)。この場合、ローカルノードのストレージセット管理部 5 0 5 は、このボリュームが完全データでないことを示すフラグをストレージセット構造情報に含まれる不完全ノードについてのプロパティから取り除く (S 1 2 7)。生成データであることを示すフラグが付されたアクセス管理情報がある場合、まだ逐次書き込みが終了していない (S 1 2 5 : N o)。この場合、一旦処理を終了し、S 1 2 5 を繰り返す。S 1 2 5 の判定が Y e s となるまで、又は、一定回数 S 1 2 5 を繰り返すまで、S 1 2 5 は繰り返されることとしてもよい。

#### 【 0 2 4 8 】

図 3 1 は、それぞれ向かって左から順に、ノード A にアクセスする利用者 A、ノード E にアクセスする利用者 E 及びノード C にアクセスする利用者 C が、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリュームを示す表である。以下、図 3 1 を用いてデータの複製又は再生を逐次に行う場合について具体的に説明する。この図 3 1 は、ノード C にアクセスする利用者が追加された際の状態を示す。なお、以下の説明においてデータは 4 つのボリュームに分割されて分散して格納されると仮定する。また、以下の説明において、利用者 C がアクセスするノード C をローカルノードという。

#### 【 0 2 4 9 】

まず、図 3 1 に示すように、利用者 C が使用するストレージセット（つまり、ノード C から見たストレージセット）は、ノード B、C、D 及び E である。ノード D 以外のノードには既にボリュームが格納されている。不足しているボリュームはボリューム d である。しかし、ノード B、C 及び E に格納されているボリューム b、a 及び c から、データを復元することは可能であるため、ノード D へのボリューム

d の書き込みを即時に行う事は不要である。

【 0 2 5 0 】

そこで、ノードDのストレージセット管理部505は、アクセス管理テーブル510から該当するストレージセット番号を持つアクセス管理情報を取得し、そのアクセス管理情報において、そのボリュームdへのアクセスに関するプロパティに、アクセスされるデータが、ストレージSから読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付す。

【 0 2 5 1 】

利用者Cからデータの読み出し要求を受信した場合、ローカルノードの制御装置Cは、ノードB、C及びEからそれぞれボリュームb、a及びcを受信し、それらのボリュームから冗長化を利用してデータを再現して利用者に転送する。その際に、ローカルノードの制御装置Cはボリュームdを作成し、ノードDのストレージSにそのボリュームdを逐次に格納させる。

【 0 2 5 2 】

図32に、制御装置の配置方法の変形例を示す。上記説明において、制御装置Cは各ノードに備えられるとした。しかし、制御装置Cは、利用者の端末に備えられる事としても良い。この場合、利用者の端末が、データを複数のボリュームに分割し、各ボリュームの格納先となるストレージを選択してデータを書き込む。さらに、利用者の端末が、複数のボリュームを各ストレージから読み出してデータを復元する。この場合でも、上記の広域分散ストレージシステムと同様の効果を得ることができる。図32には、ノードAを利用する利用者Aの端末が、データを3つのボリュームに分割し、ノードA、B及びGの3つのノードのストレージSに格納させる場合を示す。この場合、利用者Aの端末は、データを読み出す際には、ノードA、B及びGから3つのボリュームを読み出してデータを復元する。なお、この場合でも、上記の変形例に対応して様々な変形例が考え得る。

【 0 2 5 3 】

上記において説明した制御装置C内の制御部5は、コンピュータを用いて構成することができる。コンピュータは、少なくとも、CPUとそのCPUに接続されたメモリを備え、さらに、外部記憶装置、媒体駆動装置を備えてもよい。それ

らはバスにより互いに接続されている。

【 0 2 5 4 】

メモリは、例えば、ROM (Read Only Memory)、RAM (Random Access Memory) 等であり、処理に用いられるプログラムとデータを格納する。制御部 5 を構成する各部及び各テーブルは、コンピュータのメモリの特定のプログラムコードセグメントにプログラムとして格納される。なお、制御装置 C によって行われる処理は、図を用いて既に説明した。

【 0 2 5 5 】

CPU は、メモリを利用して上述のプログラムを実行することにより、必要な処理を行う。

外部記憶装置は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。外部記憶装置は、各テーブルを実現する。さらに、上述のプログラムをコンピュータの外部記憶装置に保存しておき、必要に応じて、それらをメモリにロードして使用することもできる。

【 0 2 5 6 】

媒体駆動装置は、可搬記録媒体を駆動し、その記録内容にアクセスする。可搬記録媒体としては、メモリカード、メモリスティック、フレキシブルディスク、CD-ROM (Compact Disc Read Only Memory)、光ディスク、光磁気ディスク、DVD (Digital Versatile Disk) 等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体に上述のプログラムを格納しておき、必要に応じて、それをコンピュータのメモリにロードして使用することもできる。

【 0 2 5 7 】

また、ネットワーク I/F を介して、上述のプログラムを、ネットワーク I/F を介して、プログラムをダウンロードすることとしてもよい。

以上、本発明の実施形態について説明したが、本発明は上述した実施形態及び変形例に限定されるものではなく、その他の様々な変更が可能である。

【 0 2 5 8 】

(付記 1) コンピュータが、データを冗長化して複数のボリュームに分割し、

各ボリュームを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを特徴とするデータ格納方法。

#### 【 0 2 5 9 】

(付記 2) 前記評価値の算出において、前記書き込みを依頼するノードから各ストレージまでのホップ数をさらに用いる、

ことを特徴とする付記 1 記載のデータ格納方法。

#### 【 0 2 6 0 】

(付記 3) 前記システムの利用者に対して、前記ストレージセットを仮想的な 1 つのストレージとして提供する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

#### 【 0 2 6 1 】

(付記 4) 前記データを前記ストレージセットから読み込む際には、前記ストレージセットに書きこまれた前記複数のボリュームのうち冗長化部分を含まないボリュームを各ストレージから読み出し、

前記読み出されたボリュームを用いて前記データを復元する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

#### 【 0 2 6 2 】

(付記 5) 前記データを読み込む際には、各ストレージについて、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、

前記利用優先度に基づいて、冗長化部分を含まないボリュームとして、前記複数のボリュームのうちいずれのボリュームを各ストレージから読み出すか決定する、

ことを更に含むことを特徴とする付記 3 に記載のデータ格納方法。

#### 【 0 2 6 3 】

(付記 6) 前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第 1 のボリウムの複製を格納する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 6 4 】

(付記 7) 前記第 1 のボリウムの複製を作成する際に、前記評価値に基づいて、前記第 1 のボリウムを格納するストレージから前記第 1 のボリウムを複写するのか、前記複数のボリウムのうちの前記第 1 のボリウム以外のボリウムから冗長を利用して前記第 1 のボリウムを再現するのか、2 つの作成方法のうちのいずれかを選択する、

ことを更に含むことを特徴とする付記 6 に記載のデータ格納方法。

【 0 2 6 5 】

(付記 8) 前記評価値に基づいて、前記ストレージセットとして選択されなかったストレージの中から前記ボリウムの複製を格納するストレージを選択する、

ことを更に含むことを特徴とする付記 6 に記載のデータ格納方法。

【 0 2 6 6 】

(付記 9) 同一のボリウムを格納するべき複数のストレージに対して、マルチキャストでボリウムを書き込む、

ことを更に含むことを特徴とする付記 6 に記載のデータ格納方法。

【 0 2 6 7 】

(付記 1 0) 前記第 1 のボリウムの複製をストレージに書き込む際に、多数回に分けて書き込み処理を行う、

ことを特徴とする付記 6 に記載のデータ格納方法。

【 0 2 6 8 】

(付記 1 1) 前記ストレージセットのうちの第 1 のストレージに障害が発生した場合、前記ストレージセットのうちの他のストレージへの書き込みを制限する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 6 9 】

(付記 1 2) 前記ストレージセットのうち第 3 のストレージに障害が発生した場合、前記評価値に基づいて、前記ストレージセットとして選択されているストレージ以外の第 4 のストレージを、前記第 3 のストレージの代わりに選択する

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 7 0 】

(付記 1 3) 前記ストレージセットの選択後、一定のタイミングで、各ノードにおけるストレージセットを再選択し、

再選択の結果、どのノードからも利用されていないボリュームがあった場合、該ボリュームをストレージから削除する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 7 1 】

(付記 1 4) 前記一定のタイミングとは、前回の選択から一定期間後、又はボリュームの状態が変更される毎である、

ことを特徴とする付記 1 3 に記載のデータ格納方法。

【 0 2 7 2 】

(付記 1 5) 前記データを読み込んだ後に、前記データを一定期間、任意の 1 つのストレージ内に一時格納し、

前記一定期間内にデータの読み出しを行う際には、一時格納されたデータを前記 1 つのストレージから読み出す、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 7 3 】

(付記 1 6) 一定期間内に書き込み要求されたデータを一時格納領域に一時格納し、

前記一定期間経過後に前記一時格納領域からデータを取り出し、

該データを複数のボリュームに分割し、

該複数のボリュームを前記ストレージセットに書き込む、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 7 4 】

(付記 1 7) 前記一時格納したデータを含むデータに対し、読み出し又は書き込みを行う際に、前記一時格納したデータを含まない部分のデータについてのみ、読み出し又は書き込むを行う、

ことを更に含むことを特徴とする付記 1 5 又は 1 6 に記載のデータ格納方法。

【 0 2 7 5 】

(付記 1 8) 前記ストレージセットに前記複数のボリュームを書き込む際に、前記書き込みを依頼するノードは、書き込みが終了するまで前記ストレージセットへの書き込み処理を禁止する、

ことを更に含むことを特徴とする付記 1 に記載のデータ格納方法。

【 0 2 7 6 】

(付記 1 9) 同一のボリュームを格納すべき複数のストレージの中から、1 つのストレージを代表ストレージとして決定することを更に含み、

前記複数のストレージへの書き込み処理の禁止において、

前記代表ストレージへの書き込み処理の禁止は、前記書き込みを依頼するノードによって行われ、

前記代表ストレージ以外のストレージへの書き込み処理の禁止は、前記代表ストレージによって行われる、

ことを特徴とする付記 1 8 に記載のデータ格納方法。

【 0 2 7 7 】

(付記 2 0) 前記代表ストレージは、原本となるボリュームを格納すべきストレージである、

ことを特徴とする付記 1 9 に記載のデータ格納方法。

【 0 2 7 8 】

(付記 2 1) ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリュームに分割し、各ボリュームを複数のストレージに分散して格納する制御をコンピュータに行われるコンピュータ・プログラムであって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望

ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせることを特徴とするコンピュータ・プログラム。

#### 【 0 2 7 9 】

(付記 2 2) ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリュームに分割し、各ボリュームを複数のストレージに分散して格納する制御をコンピュータに行わせるプログラムを記録した記録媒体であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせるプログラムを記録した記録媒体

。

#### 【 0 2 8 0 】

(付記 2 3) ネットワークを介して分散配置されたストレージを備えるシステムに使用される、データを冗長化して複数のボリュームに分割し、各ボリュームを複数のストレージに分散して格納する制御を行う制御装置であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する経路管理手段と、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択するストレージセット管理手段と、

を備えることを特徴とする制御装置。

#### 【 0 2 8 1 】

#### 【発明の効果】



以上詳細に説明したように、本発明によれば、データを冗長化して複数のボリュームに分割し、各ボリュームを複数のストレージに分散して格納することによりデータのセキュリティを向上させつつも、さらに、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離を考慮して、そのノードから見て最適なストレージを選択することにより、回線効率と災害時のデータの安全性も向上させることが可能となる。

【図面の簡単な説明】

【図 1】

広域分散ストレージシステムの構成図である。

【図 2】

制御装置の構成図である。

【図 3】

制御装置の詳細構成図である。

【図 4】

具体的な広域分散ストレージシステムの構成例を示す図である。

【図 5】

経路評価テーブルの一例を示す図である。

【図 6】

ストレージ評価テーブルの一例を示す図である。

【図 7】

ストレージセット管理テーブルの一例を示す図である。

【図 8】

アクセス管理テーブルの一例を示す図である。

【図 9】

ローカルボリューム管理テーブル一例を示す図である。

【図 10】

広域分散ストレージシステムにおけるデータの流れを説明する図である。

【図 11】

利用優先度及び評価値の計算処理を示すフローチャートである。

【図 1 2】

データ復元時にボリウムを読み出すべきストレージの決定方法を説明する図である。

【図 1 3】

ストレージセット管理テーブルの更新処理を示すフローチャートである。

【図 1 4】

冗長化したデータの分散書き込み先となるストレージの決定方法を説明する図である。

【図 1 5】

ノードへの利用者の追加処理を示すフローチャートである。

【図 1 6】

冗長化したデータの複製を格納するストレージの決定方法を説明する図である。

【図 1 7】

利用可能な複数のストレージから最適なストレージセットを選択する処理を説明する図である。

【図 1 8】

ロック処理を示すフローチャート（その 1）である。

【図 1 9】

ロック処理を示すフローチャート（その 2）である。

【図 2 0】

書き込み処理を示すフローチャート（その 1）である。

【図 2 1】

書き込み処理を示すフローチャート（その 2）である。

【図 2 2】

ストレージ内のデータ更新の際のロック処理を説明する図である。

【図 2 3】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その 1）である。

【図 2 4】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その 2）である。

【図 2 5】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その 3）である。

【図 2 6】

ボリウムの複製の作成方法を選択する処理を説明する図である。

【図 2 7】

一部のストレージに障害が発生した場合に、残りのストレージの中から最適なストレージを選択しなおす処理を説明する図である。

【図 2 8】

ストレージが障害から復旧した際にボリウムの複製の作成方法を選択する処理を説明する図である。

【図 2 9】

不要となったボリウムを削除する処理を説明する図である。

【図 3 0】

逐次にデータを書き込み又は再生する処理を示すフローチャートである。

【図 3 1】

データの複製の作成又は再生を逐次に行う場合の処理を説明する図である。

【図 3 2】

制御装置の機能を利用者端末が備える場合を説明する図である。

【図 3 3】

従来の技術に係わる広域分散ストレージシステムの構成図である。

【符号の説明】

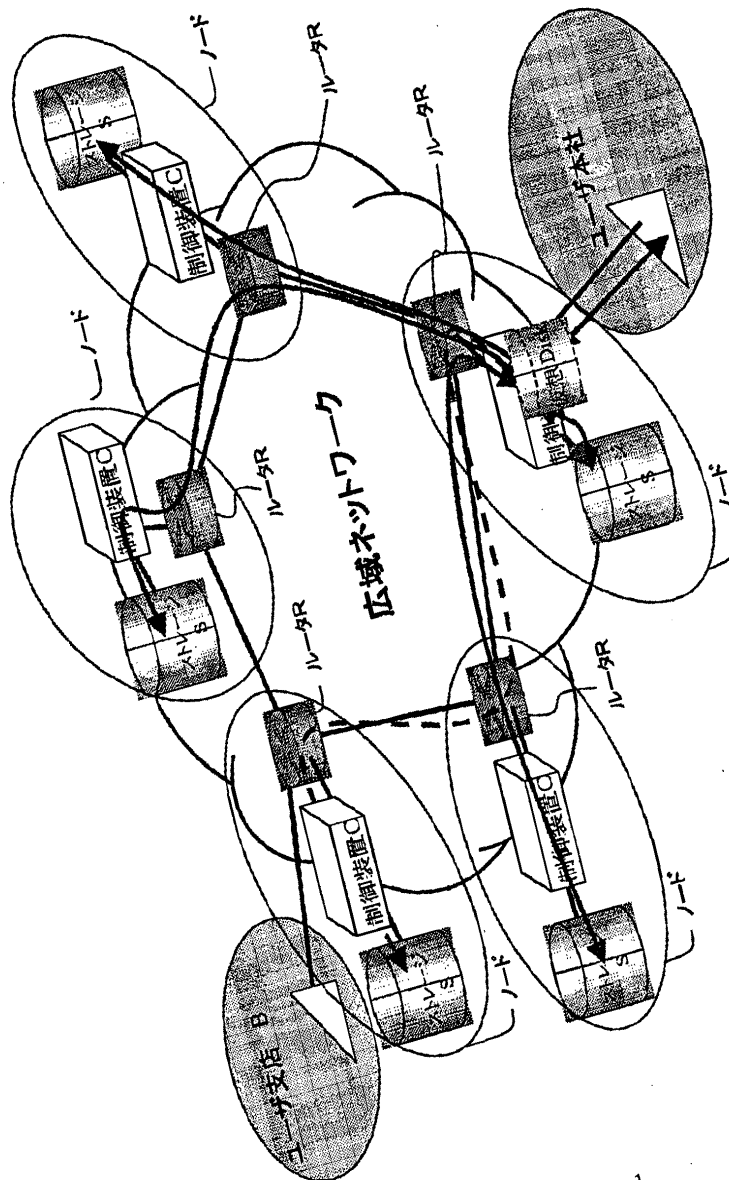
- 1 ユーザインタフェース（受信側）
- 2 ユーザインタフェース（送信側）
- 3 データ変換部
- 4 パケット生成部

- 5 制御部
- 6 データ組立部
- 7、3 0 1 パケット解析部
- 8 ストレージインタフェース
- 9 ネットワークインタフェース（送信側）
- 1 0 ネットワークインタフェース（受信側）
- 3 0 2 データ分割部
- 3 0 3、6 0 3 パリティ計算部
- 4 0 1 データ管理情報付加部
- 4 0 2 制御／経路情報付加部
- 4 0 3 データ転送部
- 4 0 4 転送パケット構築部
- 5 0 1 ストレージ制御部
- 5 0 2 制御パケット生成部
- 5 0 3 ネットワーク制御部
- 5 0 4 経路管理部
- 5 0 5 ストレージセット管理部
- 5 0 6 ローカルボリューム管理部
- 5 0 7 経路評価テーブル
- 5 0 8 ストレージ評価テーブル
- 5 0 9 ストレージセット管理テーブル
- 5 1 0 アクセス管理テーブル
- 5 1 1 ローカルボリューム管理テーブル
- 6 0 1 パケット構築部
- 6 0 2 データ組立部
- C 制御装置
- R ルータ
- S ストレージ

【書類名】  
【図1】

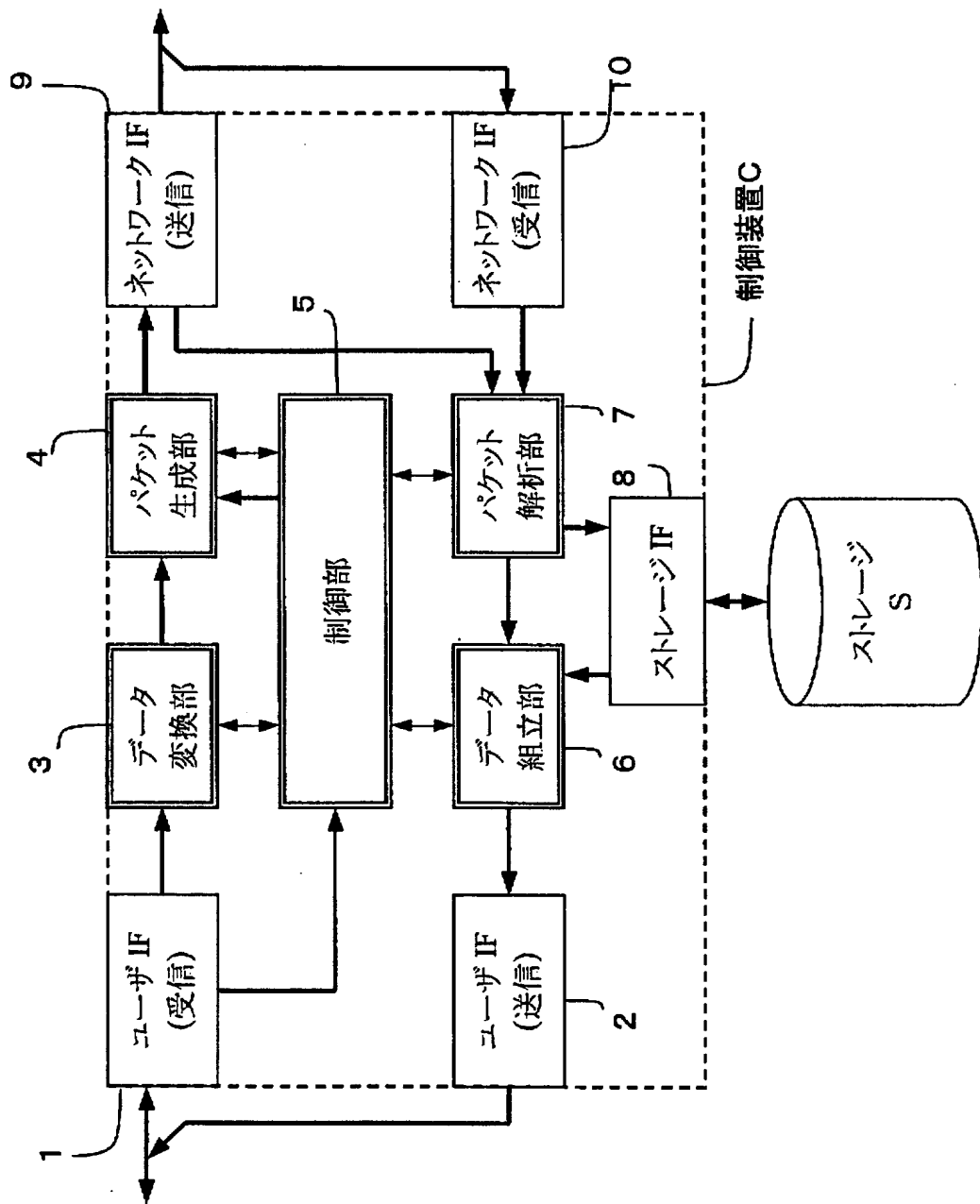
図面

広域分散ストレージシステムの構成図



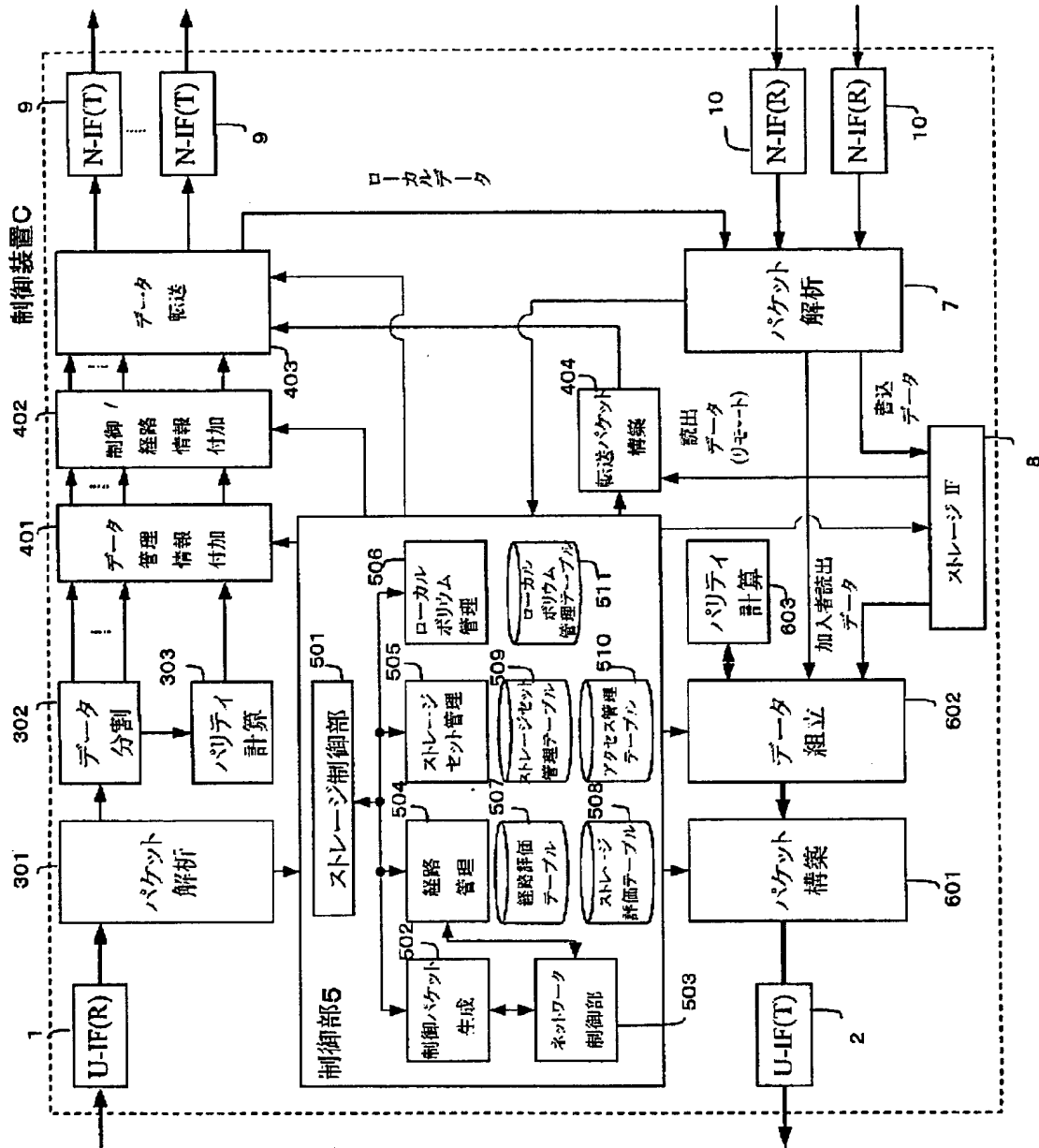
【図 2】

制御装置の構成図



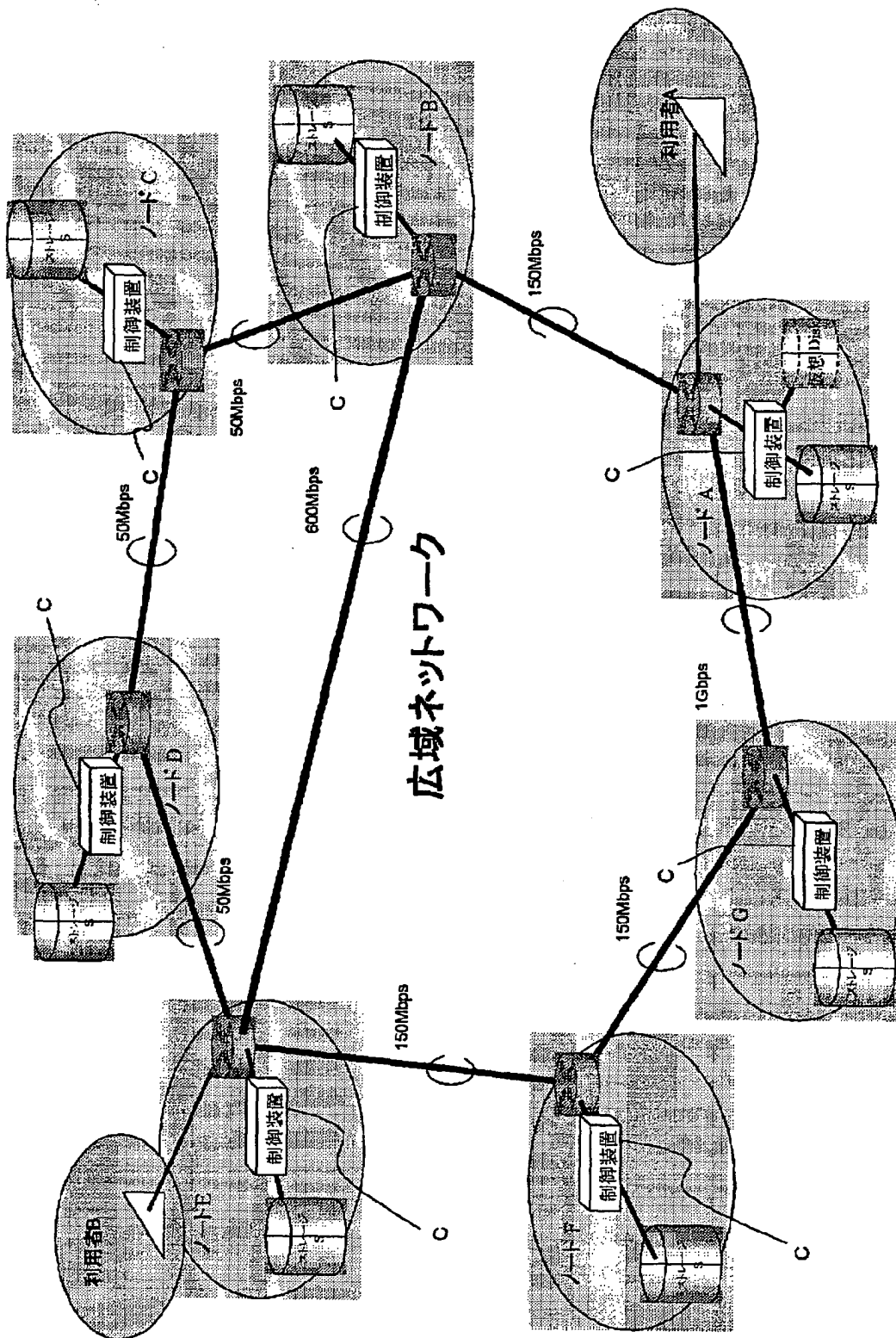
【図 3】

制御装置の詳細構成図



【図4】

具体的な広域分散ストレージシステムの構成例を示す図





【図 5】

## 経路評価テーブルの一例を示す図

経路評価テーブル

区間	帯域幅	コスト	ディスタンス	ストレージの利用 優先度
local	-	-	-	$\infty$
A-B	150	100	80	11
B-C	50	50	150	17
C-D	50	100	200	21
D-E	50	50	150	17
E-F	150	100	100	13
F-G	150	100	80	11
A-G	1000	200	10	11
B-E	600	300	200	24

【図 6】

## ストレージ評価テーブルの一例を示す図

ノード	経路	ストレージ 評価値	ホップ数
A	local	$\infty$	0
B	A-B	11.0	1
C	A-B-C	9.8	2
D	A-B-C-D	8.8	3
	A-B-E-D	9.6	3
E	A-B-E	11.5	2
F	A-G-F	8.3	2
G	A-G	11.0	1

【図 7】

ストレージセット管理テーブルの一例を示す図

ストレージセット 番号	プロパティ																							
	全体			ノード A			ノード B			ノード C			ノード D			ノード E			ノード F			ノード G		
	読 出	書 込	状 態	分 割	原 本	使 用	分 割	原 本	使 用	分 割	原 本	使 用	分 割	原 本	使 用	分 割	原 本	使 用	分 割	原 本	使 用	分 割	原 本	使 用
00000001	3	4	G	1	O	RW	3	O	RW	2	O	-W				0	O	RW				2	C	R
00000002	3	3	G	2	O	Rw				1	O	RW	0	O	RW									
00000003	4	5	R	0	O	RW	1	O	RW	4	O	RW	3	O	RW	2	O	-W						
00000004	2	3	G			RW	2	O		0	O	-W				1	O	RW						

【図 8】

アクセス管理テーブルの一例を示す図

ストレージセット番号	000010001
ストレージセット アクセス番号	プロパティ
00000001	RWLO
00000002	RWLO
00000003	-W-O
00000004	-W-O
00000005	---O
00000006	-W-O

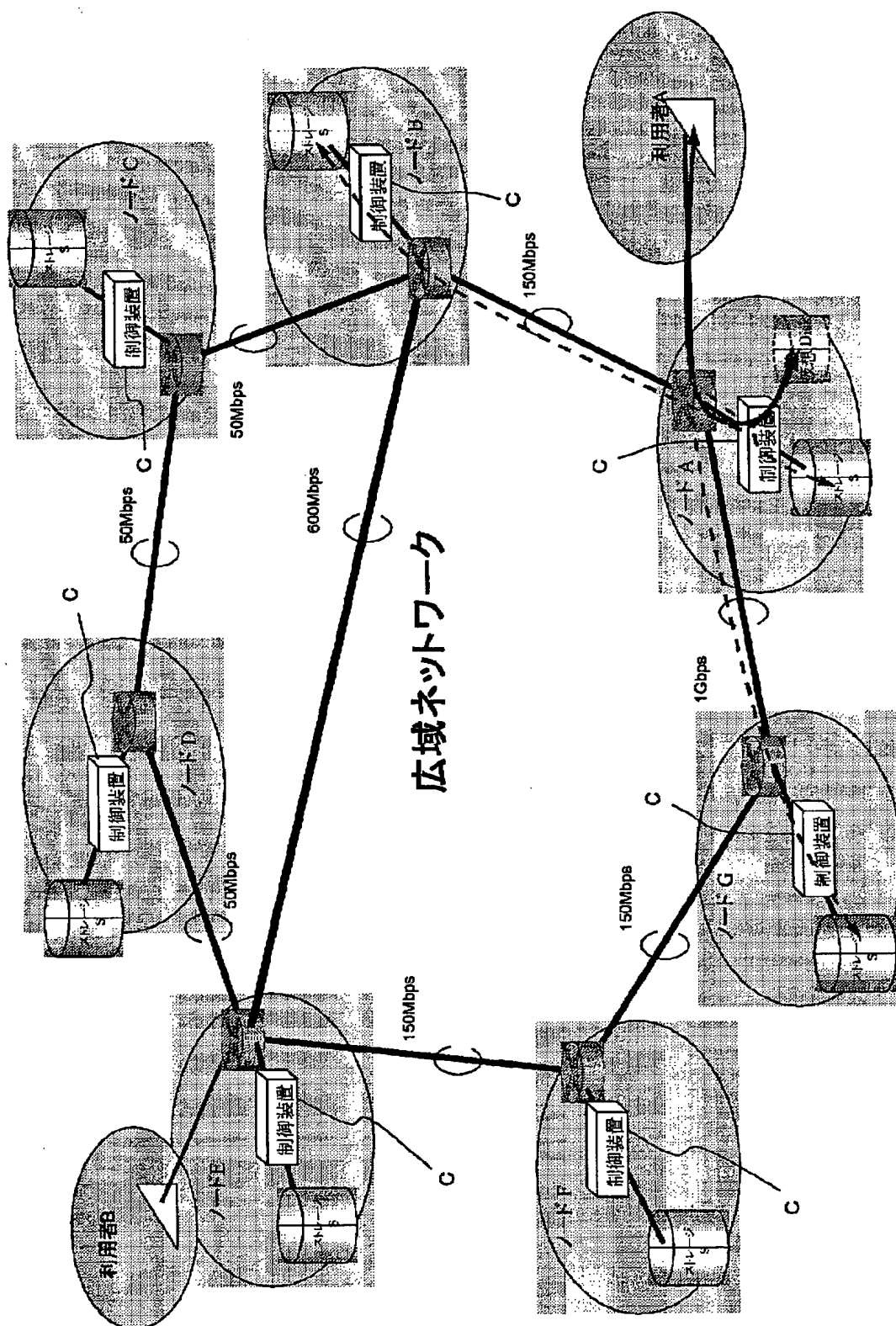
【図 9】

## ローカルポリウム管理テーブルの一例を示す図

ローカルストレージ		ストレージセット管理情報	
ストレージ アクセス番号	プロパティ	ストレージセット 管理番号	ストレージセット アクセス番号
00000001	RWL	00010021	00040001
00000002	RW-	02003001	00040004
00000003	RW-	00000000	00000000
00000004	R-F	01002111	01000100
00000005	RW-	00010021	00001562
00000006	RW-	00010021	00001563

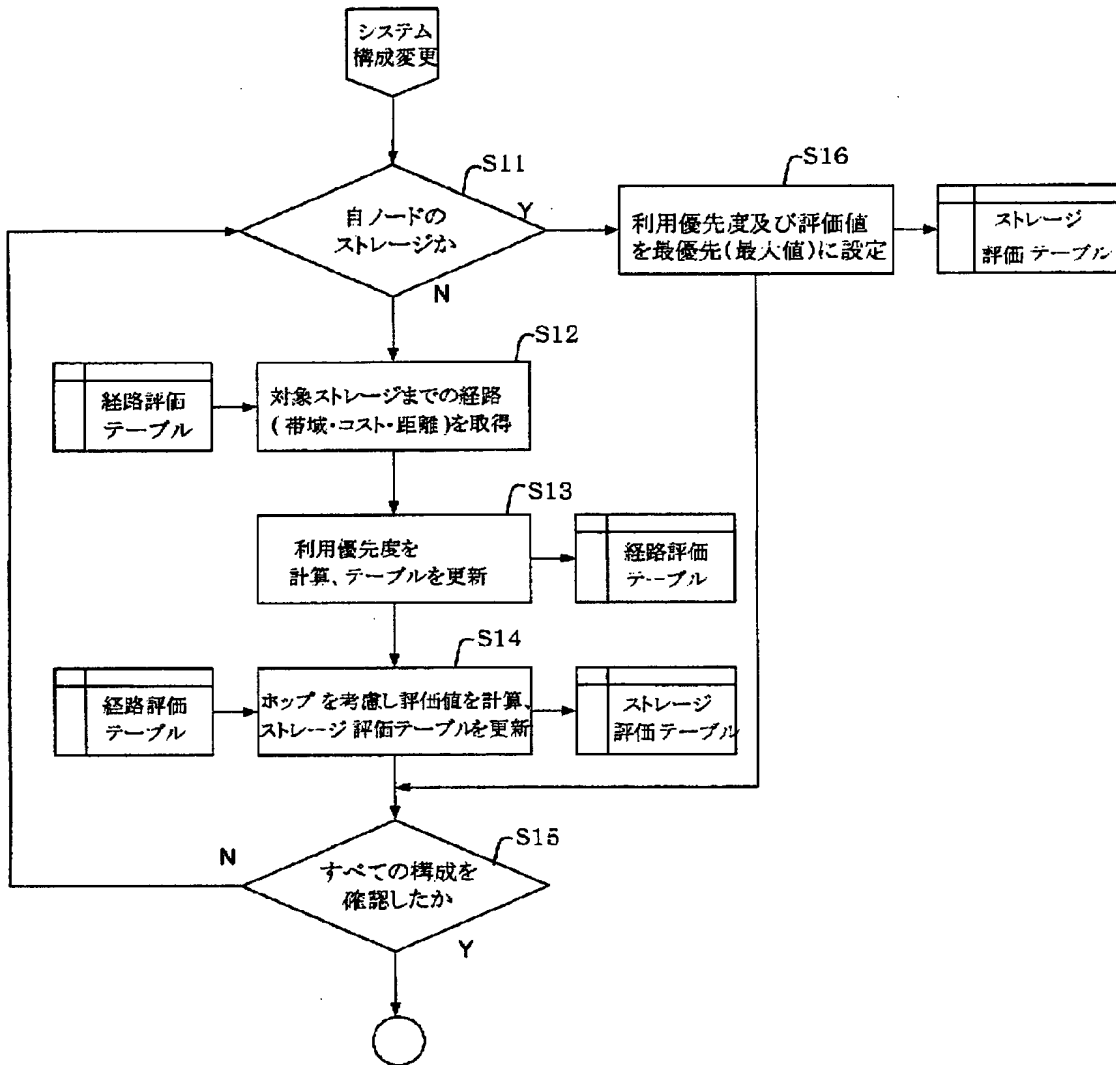
【図10】

広域分散ストレージシステムにおけるデータの流を説明する図



【図 1 1】

利用優先度の計算処理を示すフローチャート



【図 1 2】

データ復元時に、読み出すべき  
ボリウムの決定方法を説明する図

経路評価テーブル

区間	帯域幅	コスト	ディスタンス	区間の利用優先度	ストレージの利用 優先度
local	-	-	-	$\infty$	$\infty$
A-B	150	100	80	3	11
B-C	50	50	150	2	17
C-D	50	100	200	1	21
D-E	50	50	150	2	17
E-F	150	100	100	3	13
F-G	150	100	80	3	11
A-G	1000	200	10	10	11
B-E	600	300	200	4	24

(a)

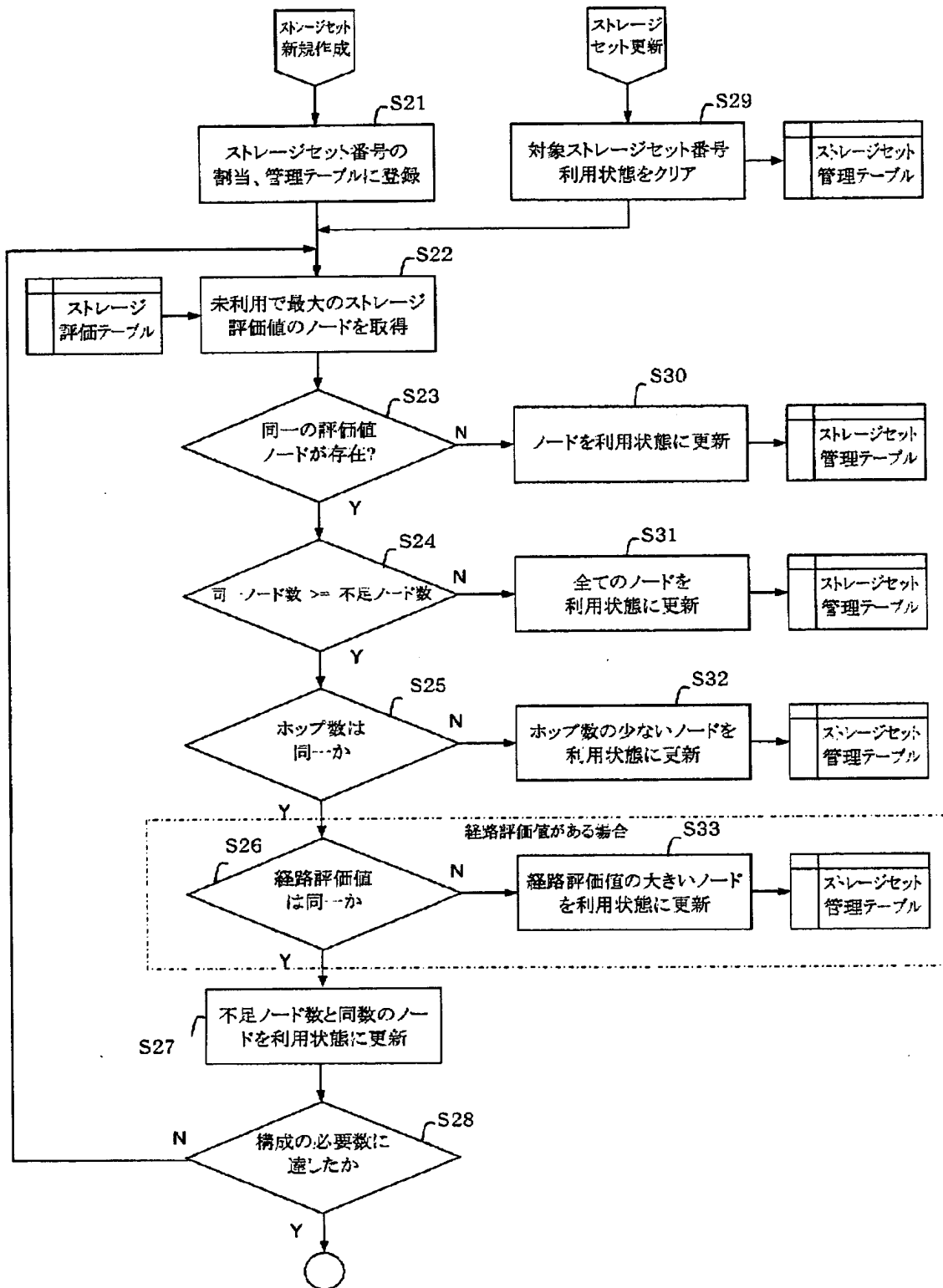
ストレージ評価テーブル

ノード	経路	経路評価値	ストレージ 評価値	ホップ数
A	local	$\infty$	$\infty$	0
B	A-B	3	11.0	1
C	A-B-C	2.5	9.8	2
D	A-B-C-D	2	8.8	3
	A-B-E-D	3	9.6	3
E	A-B-E	3.5	11.5	2
F	A-G-F	6.5	8.3	2
G	A-G	10	11.0	1

(b)

【図 13】

ストレージセット管理テーブルの更新処理を示すフローチャート



【図 14】

冗長化したデータの分散書き込み先となる  
ストレージの決定方法を説明する図

経路評価テーブル

区間	帯域[Mbps]	コスト	ディスタンス	ストレージ利用優先度
A-B	150	100	80	11
B-C	50	50	150	17
C-D	50	100	200	21
D-E	50	50	150	17
E-F	150	100	100	13
F-G	150	100	80	11
A-G	1000	200	10	11
B-E	600	300	200	24

(a)

ストレージ評価テーブル

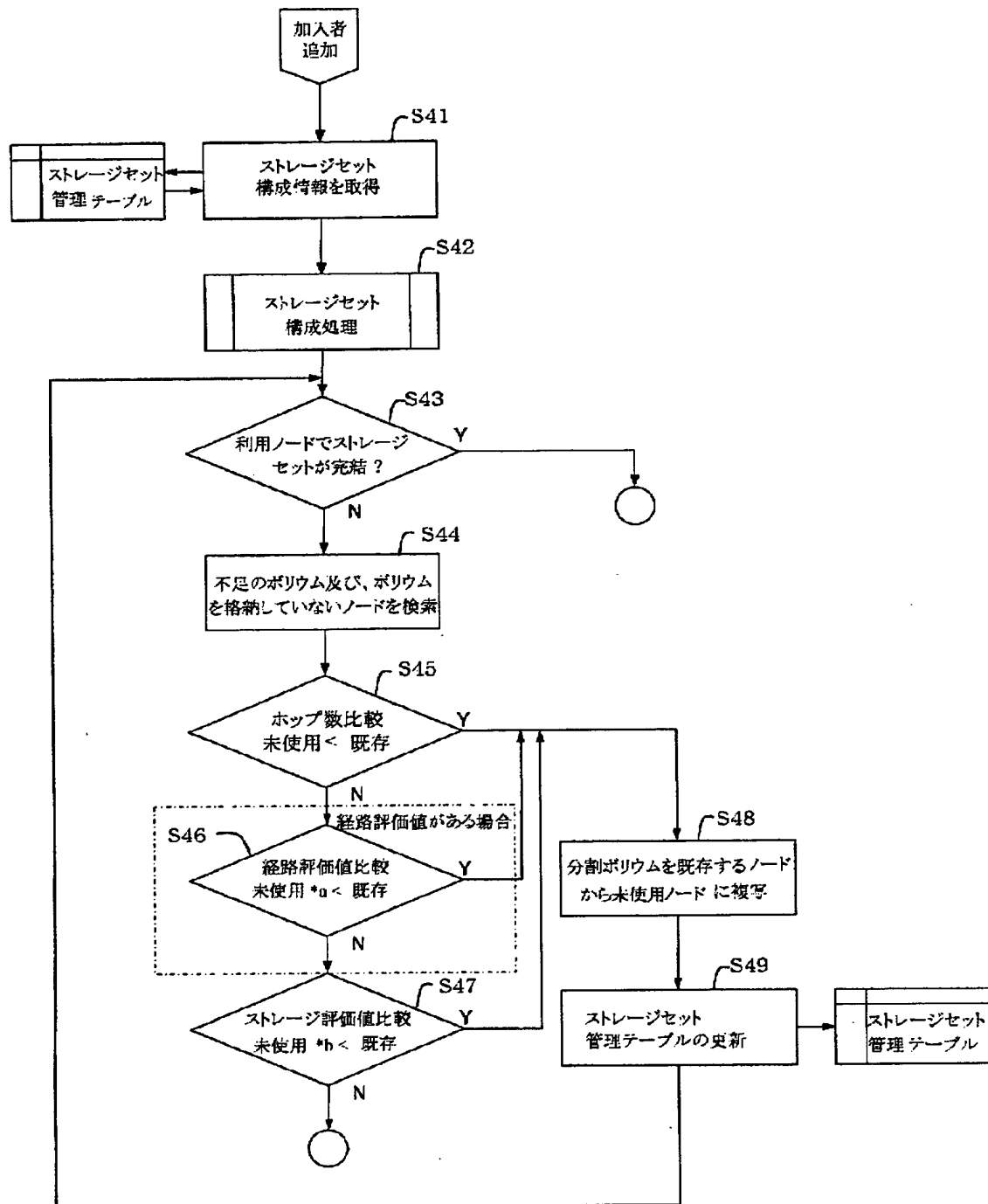
ノード	経路	式	ストレージ評価値
A			$\infty$
B	A-B	$= (A-B)$	11.0
C	A-B-C	$= \{(A-B)+0.5(B-C)\}/2$	9.8
D	A-B-E-D	$= \{(A-B)+0.5(B-E)+0.33(D-E)\}/3$	9.6
E	A-B-E	$= \{(A-B)+0.5(B-E)\}/2$	11.5
F	A-G-F	$= \{(A-G)+0.5(F-G)\}/2$	8.3
G	A-G	$= (A-G)$	11.0

(b)



【図 15】

ノードへの利用者の追加処理を示すフローチャート



【図 1 6】

冗長化したデータの複製を格納する  
ストレージの決定方法を説明する図

ノード	利用者 A			利用者 B		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	a	17.8	2	←
B	11.0	1	b	24.0	1	←
C	9.8	2		16.3	2	
D	9.6	3		17.0	1	d'
E	11.5	2	c	$\infty$	0	←
F	8.3	2		13.0	1	
G	11.0	1	d	10.8	2	←

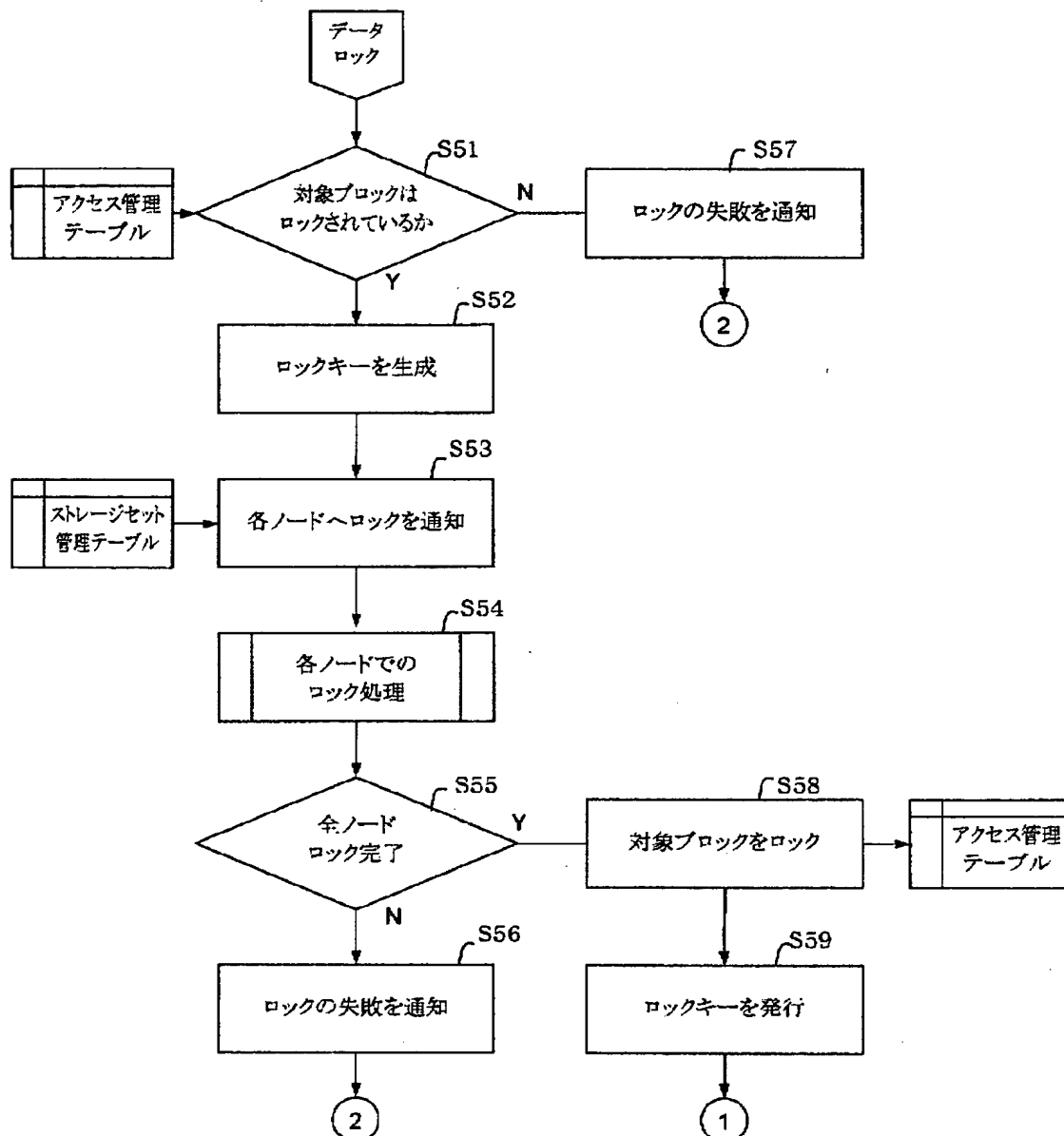
【図 1 7】

利用可能な複数のストレージから  
最適なストレージセットを選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	a	17.8	2	←	11.3	2	←
B	11.0	1	b	24.0	1	←	17.0	1	←
C	10.8	2		16.3	2		$\infty$	0	a'
D	9.1	3		17.0	1	d'	21.0	1	←
E	11.5	2	c	$\infty$	0	←	14.8	2	←
F	8.3	2	a'	13.0	1	←	10.9	3	←
G	11.0	1	d	10.8	2	←	8.4	3	←

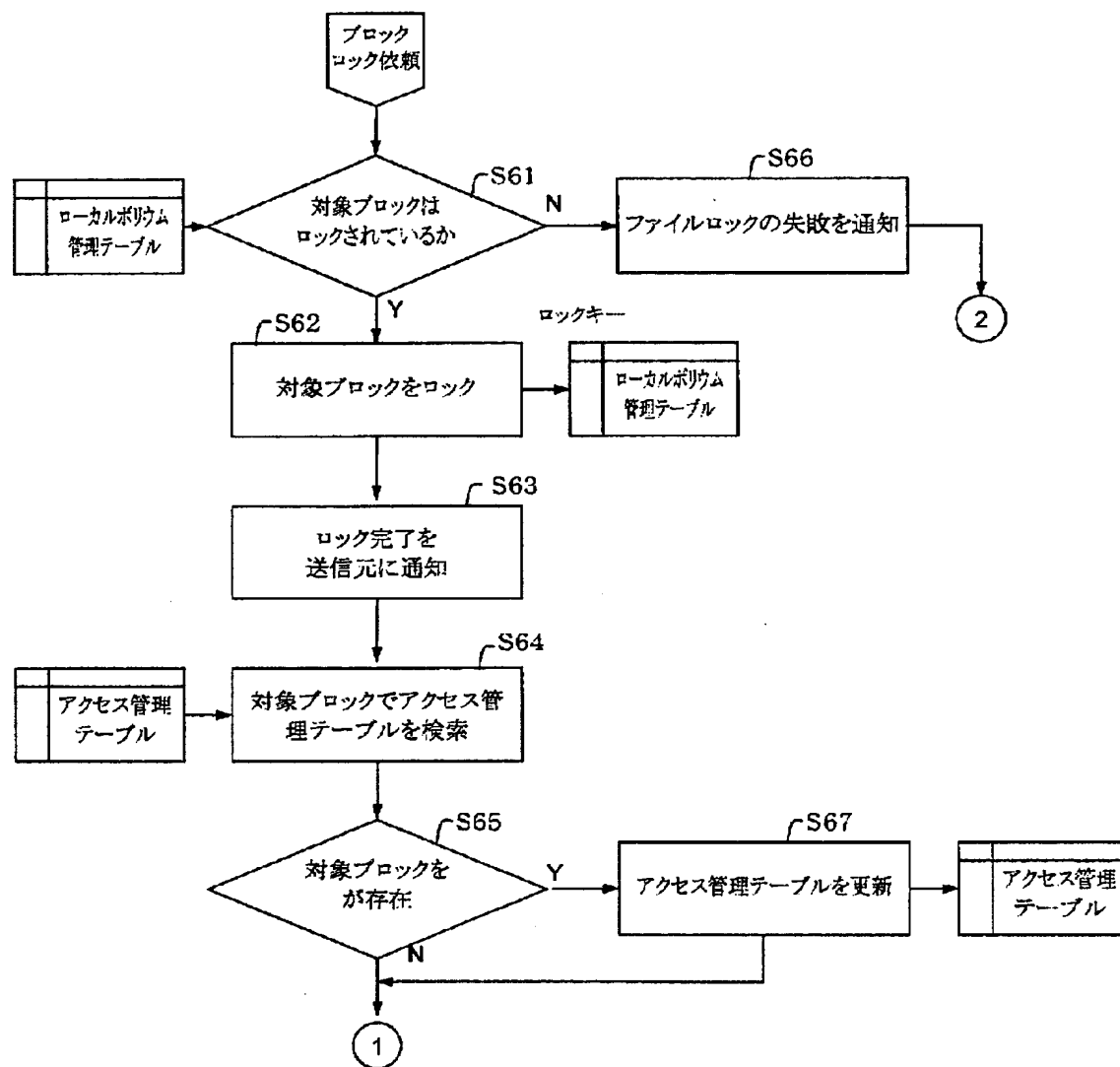
【図 1 8】

ロック処理を示すフローチャート(その1)



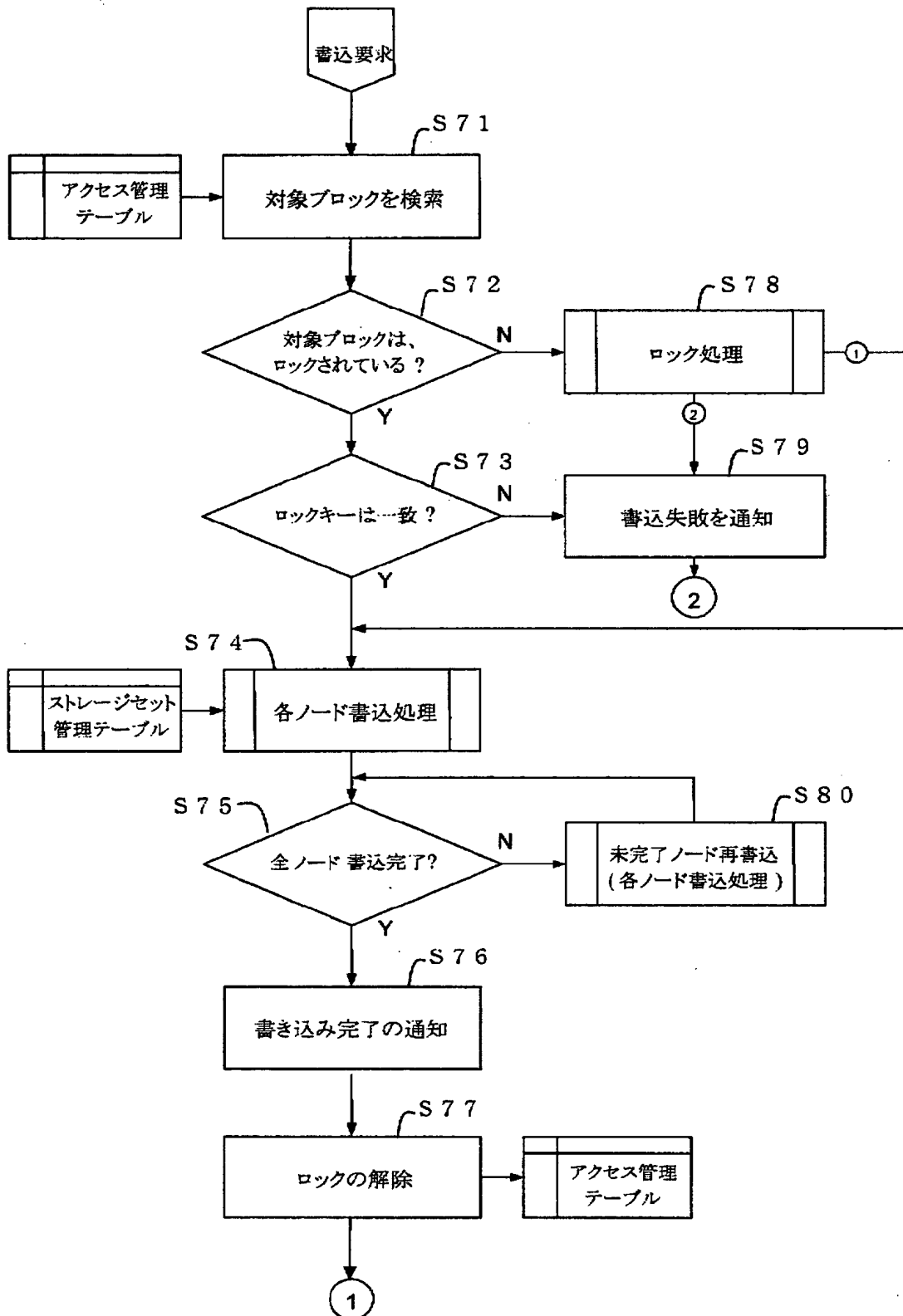
【図 1 9】

## ロック処理を示すフローチャート(その2)



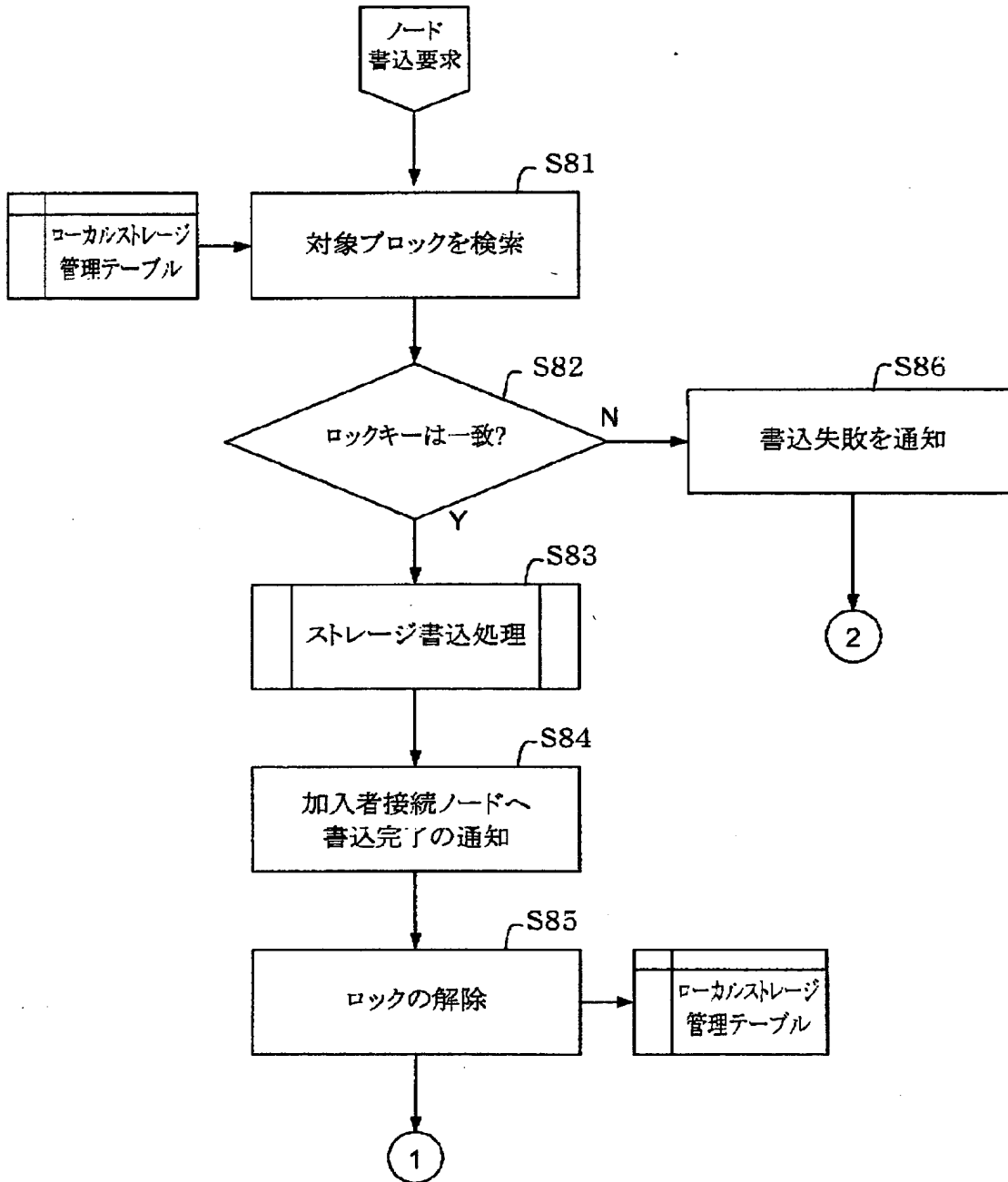
【図 2 0】

## 書き込み処理を示すフローチャート(その1)



【図 2 1】

書き込み処理を示すフローチャート(その2)



【図 2 2】

## ストレージ内のデータ更新の際のロック処理を説明する図

ストレージセット番号	000010001	
ストレージセット アクセス番号	プロパティ	ロックキー
00000001	RWLO	0111344
00000002	RWLO	1124433
00000003	-W-O	
00000004	-W-O	
00000005	---O	
00000006	-W-O	

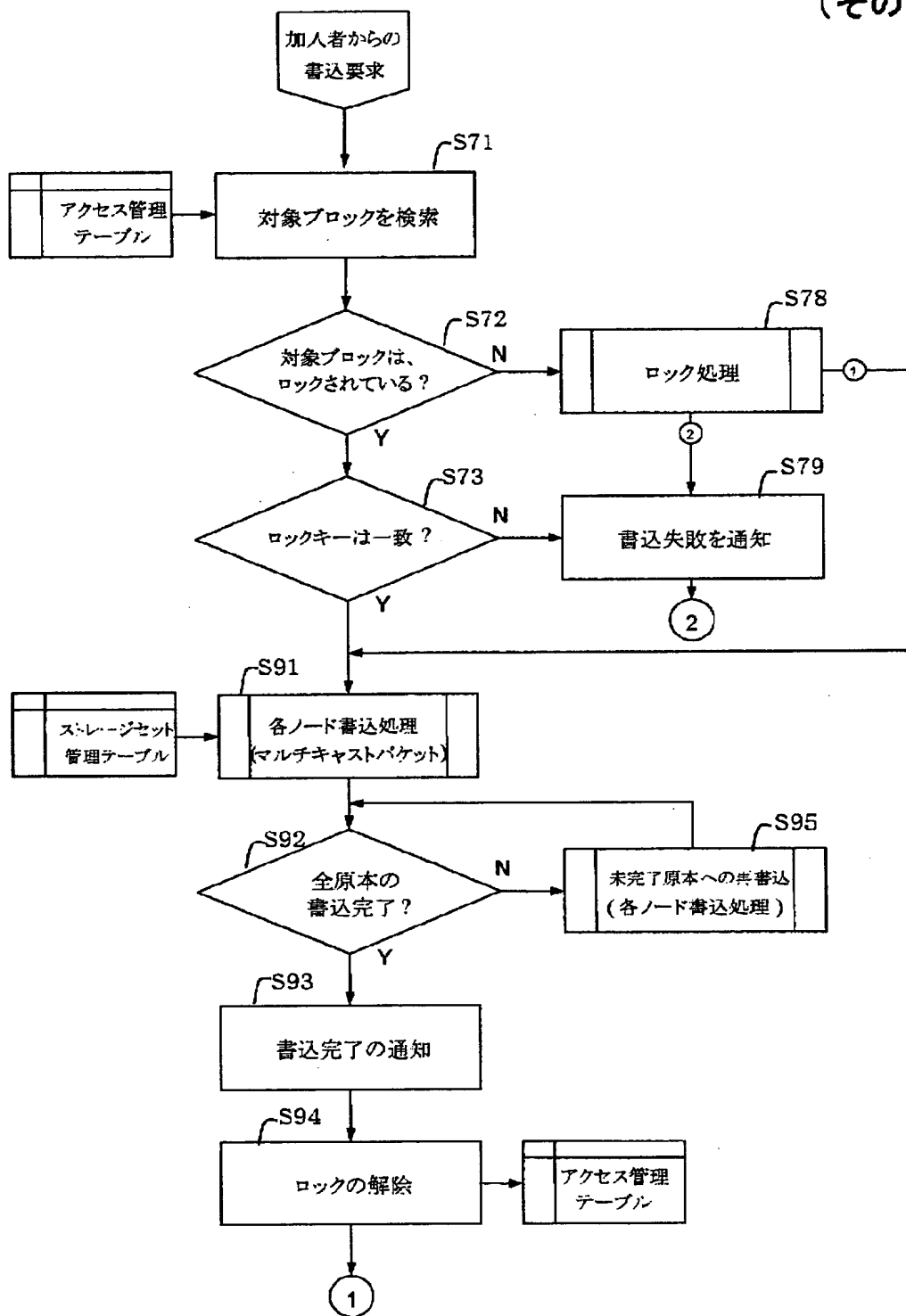
( a )

ローカルストレージ			ストレージセット管理情報	
ストレージ アクセス番 号	プロパティ	ロックキー	ストレージセット 管理番号	ストレージセット アクセス番号
00000001	RWL		00010021	00040001
00000002	RW-	001223	02003001	00040004
00000003	RW-		00000000	00000000
00000004	R-F		01002111	01000100
00000005	RW-		00010021	00001562
00000006	RW-		00010021	00001563

( b )

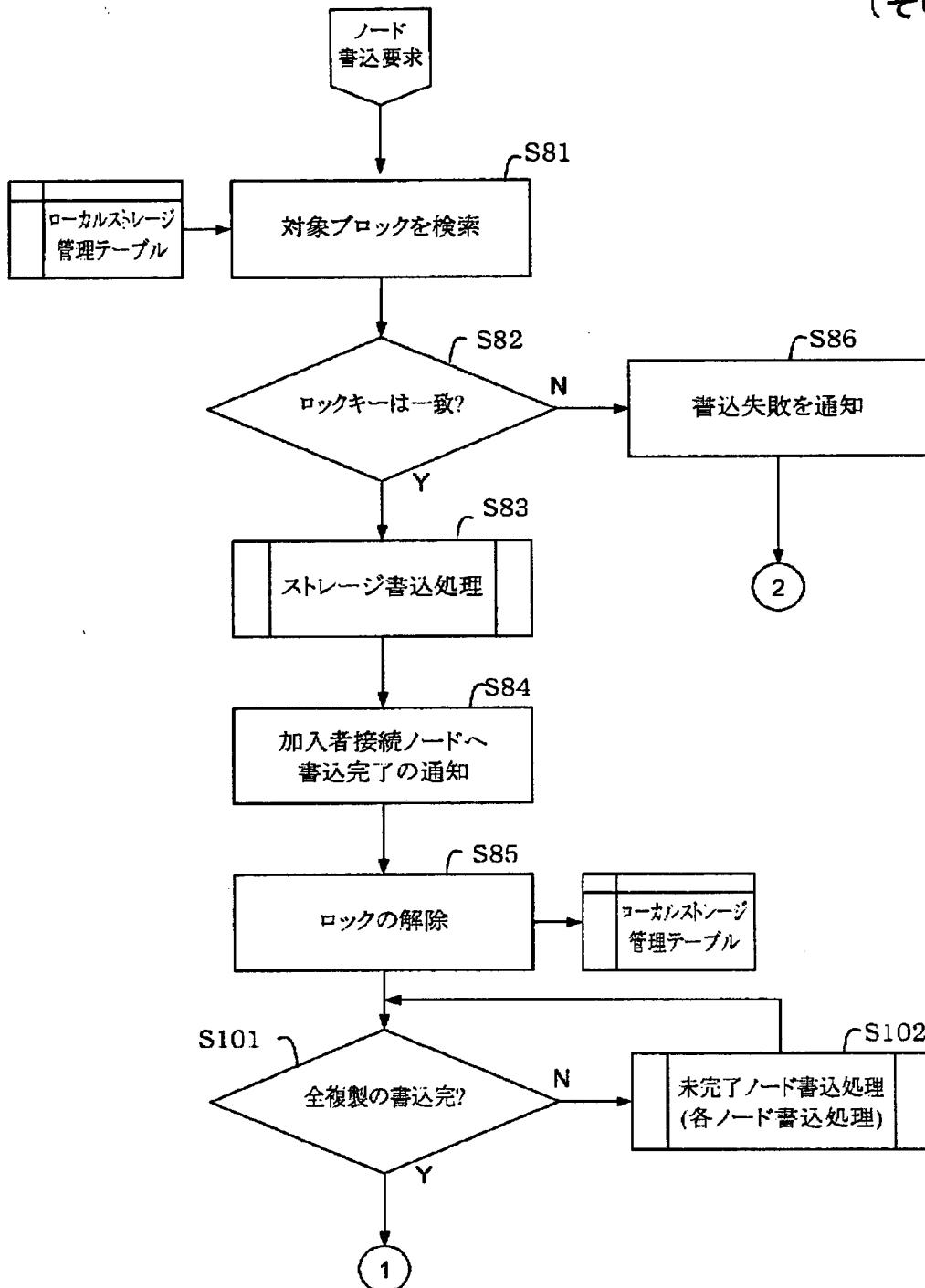


【図 23】

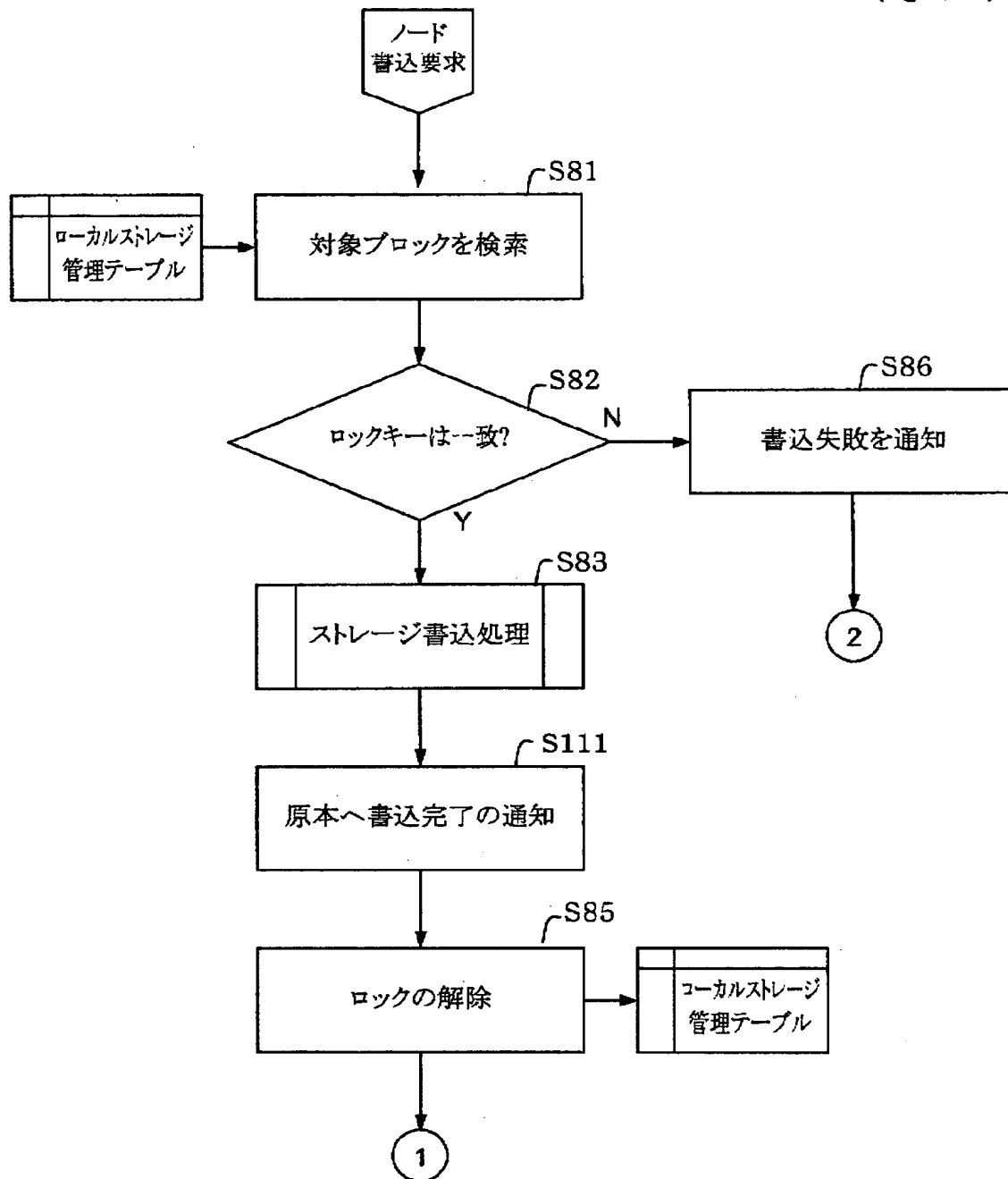
マルチキャスト packets を用いた書き込み処理を示すフローチャート  
(その1)

【図 2 4】

マルチキャストパケットを用いた書き込み処理を示すフローチャート  
(その2)



【図 2 5】

マルチキャストパケットを用いた書き込み処理を示すフローチャート  
(その3)

【図 2 6】

## ポリウムの複製の作成方法を選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	-	16.3	2	-	$\infty$	0	-
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d'

【図 2 7】

一部のストレージに障害が発生した場合に、残りのストレージの中から最適なストレージを選択しなおす処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリューム	評価値	ホップ数	ボリューム	評価値	ホップ数	ボリューム
A	$\infty$	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a'	16.3	2	a'	$\infty$	0	a'
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

(a)

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリューム	評価値	ホップ数	ボリューム	評価値	ホップ数	ボリューム
A	$\infty$	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	$\infty$	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

(b)

【図 2 8】

ストレージが障害から復旧した際に  
ポリウムの複製の作成方法を選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	-	17.8	2	-	11.3	2	-
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	$\infty$	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

【図 2 9】

不要となった複製ポリウムを削除する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	$\infty$	0	a
D	9.1	3	d	17.0	1	d	21.0	1	d
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a	13.0	1	a	10.9	3	a
G	11.0	1	d	10.8	2	d	8.4	3	d

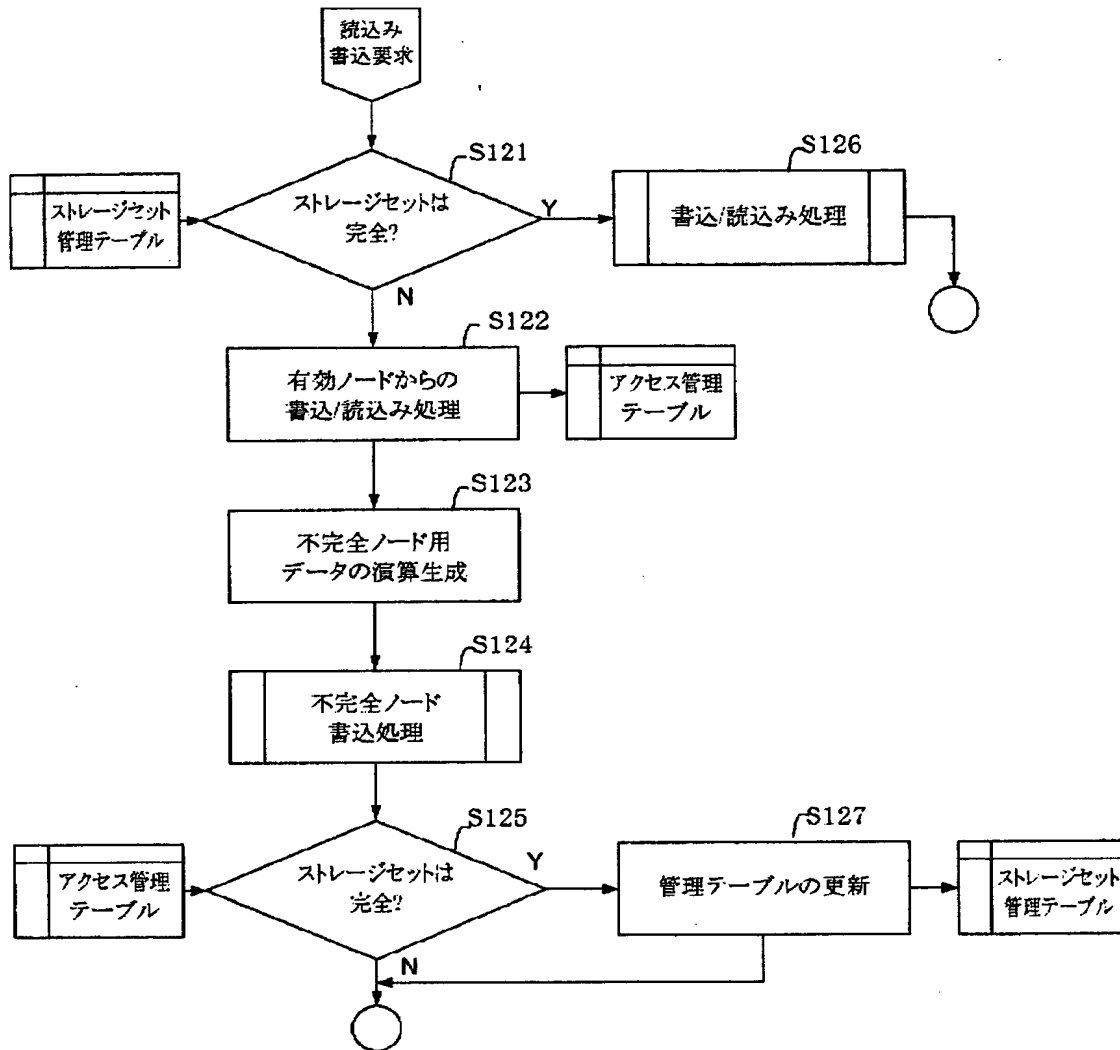
(a)

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	$\infty$	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	$\infty$	0	a
D	9.1	3	d	17.0	1	d	21.0	1	d
E	11.5	2	c	$\infty$	0	c	14.8	2	c
F	8.3	2	a	13.0	1	a	10.9	3	a
G	11.0	1	d	10.8	2	d	8.4	3	d

(b)

【図 3 0】

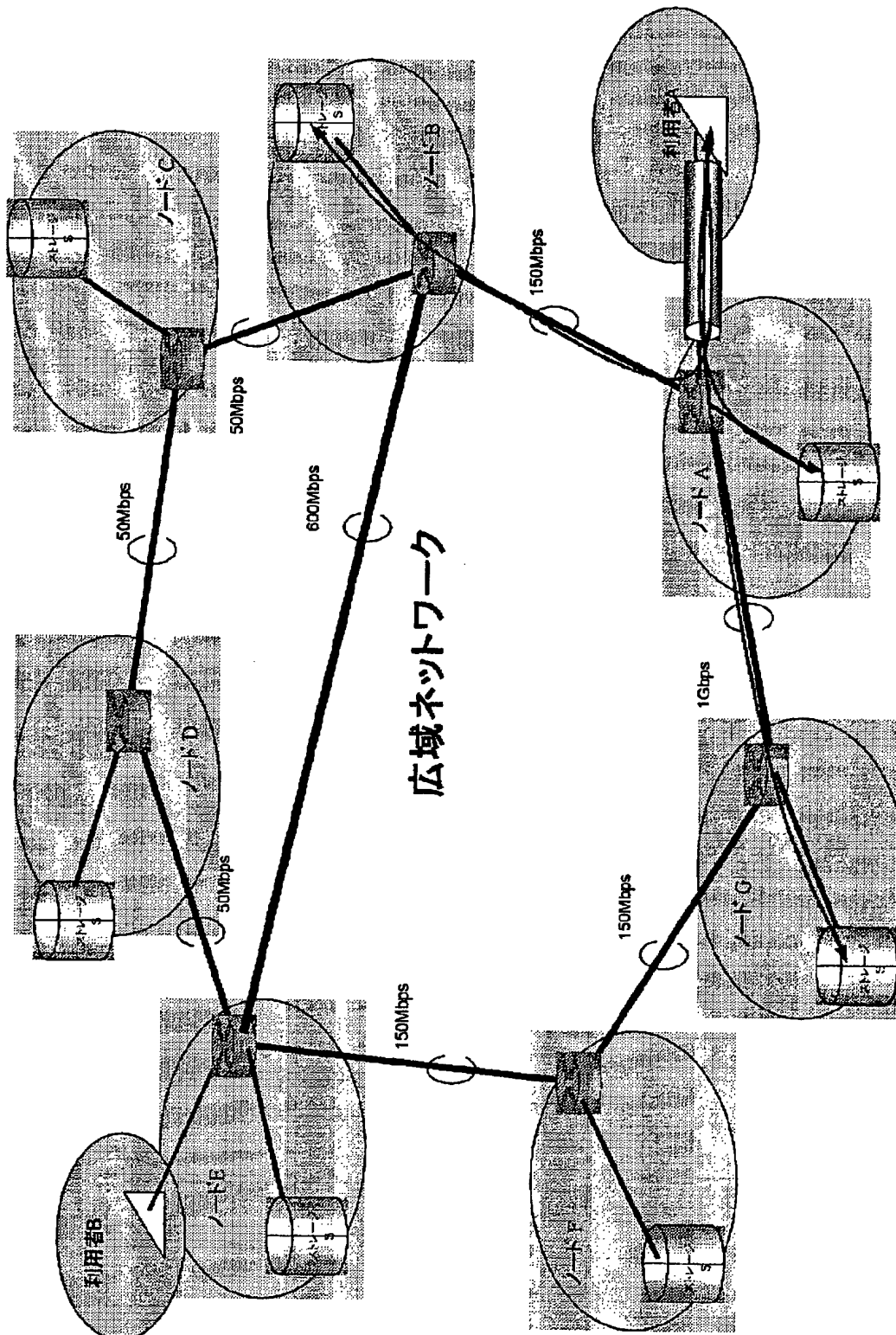
逐次にデータを書き込み又は再生する処理を示すフローチャート



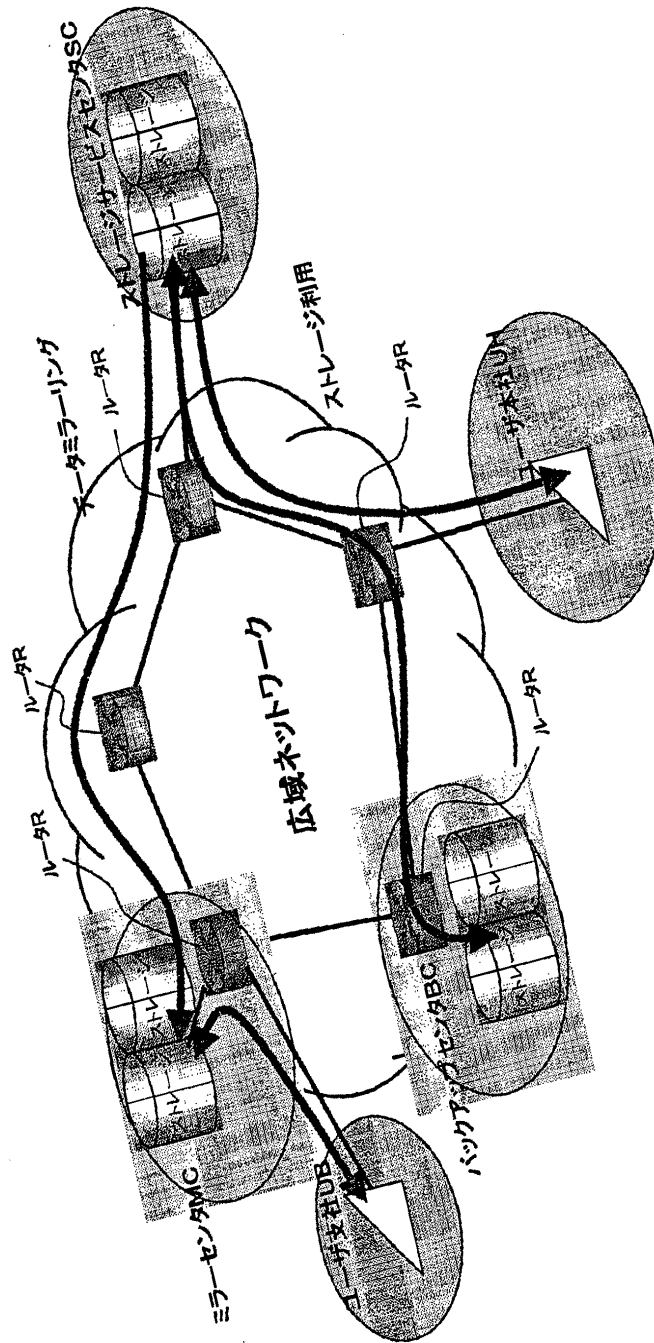


【図 3 2】

制御装置の機能を  
利用者端末が備える場合を説明する図



【図33】  
従来の技術に係わる広域分散ストレージシステムの構成図



【書類名】 要約書

【要約】

【課題】 データの冗長化に要するストレージ容量を低減しつつも、データのセキュリティを向上させ、且つ、回線を効率的に利用することが可能な R A I D を提供する。

【解決手段】 データを冗長化して複数のボリュームに分割し、各ボリュームを、ネットワークを介して分散配置された複数のストレージ S に分散して格納する制御を行う制御装置 C は、経路管理部 5 0 4 と、ストレージセット管理部 5 0 5 を備える。経路管理部 5 0 4 は、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する。ストレージセット管理部 5 0 5 は、前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する。

【選択図】 図 3